Future Proves Past: Detecting Hidden Fraud under Imperfect Enforcement Labels

C. Wei Li* Yangen Li† Yuxuan Li‡ Anand M. Vijh§

July 31, 2026

Abstract

SEC enforcement labels are selective and incomplete proxies for the latent fraud they are meant to measure. This mismatch creates two problems: a model trained on them may learn the enforcement process rather than the fraud process, and accuracy measured against them cannot reveal whether a model detects the fraud enforcement misses. We address both with a *future-proves-past* design. Because accounting distortions eventually reverse, future accounting outcomes modestly refine the training signal into soft labels; we train a convolutional neural network on multi-year accounting panels with these labels and validate it entirely outside the enforcement set. The model preserves SEC-labeled detection performance. Out of sample over 2000–2019, flagged firms complete seasoned equity offerings at nearly three times the base rate and earn large negative post-issuance abnormal returns, a long-short portfolio earns significant factor-adjusted alpha, and externally screened populations (post-scrutiny Arthur Andersen clients and bank-monitored borrowers) are systematically avoided. An identically specified network trained on the enforcement label, and benchmarks designed to predict it, show none of this profile: the training target drives the difference. The results show that selective enforcement data can be repurposed, at modest cost, to reach fraud that enforcement itself never surfaces, and they caution that detection models should be judged by validation outside the very label on which they were trained.

Keywords: financial statement fraud; hidden fraud; SEC enforcement; weak supervision; deep learning

*C. Wei Li, Department of Finance, Tippie College of Business, University of Iowa, Iowa City, IA 52242-1000. Email: wei-li@uiowa.edu

†Yangen Li, Department of Finance, Tippie College of Business, University of Iowa, Iowa City, IA 52242-1000. Email: yangen-li@uiowa.edu

‡Yuxuan Li, Department of Management Science and Engineering, School of Business, East China University of Science and Technology, Shanghai, China. Email: yuxuan.li@ecust.edu.cn

§Anand M. Vijh, Department of Finance, Tippie College of Business, University of Iowa, Iowa City, IA 52242-1000. Email: anand-vijh@uiowa.edu

1 Introduction

Financial statement fraud threatens market integrity, investor confidence, and the efficient allocation of capital (Lie, 2025). A large accounting literature has developed methods to detect fraud and earnings manipulation using accounting information, ranging from ratio-based screening models such as the M-score and F-score (Beneish, 1999; Dechow et al., 2011) to more recent machine-learning approaches that incorporate financial statements, textual disclosures, and ensemble methods (Kirkos et al., 2007; Ravisankar et al., 2011; Larcker and Zakolyukina, 2012; Brown et al., 2020; Bao et al., 2020). A common feature of this literature is the use of SEC enforcement actions, particularly Accounting and Auditing Enforcement Releases (AAERs), as the labels that stand in for actual fraud in model training and evaluation; surveying top-journal fraud research, Donelson et al. (2021) find that a majority of studies measure fraud with SEC cases alone.

However, recent evidence from Dyck et al. (2024) suggests that corporate fraud is substantially more pervasive than implied by observed SEC enforcement. Enforcement is constrained by institutional resources and legal standards, and its selection is systematic rather than random: investigations depend on factors such as geographic proximity to SEC offices (Kedia and Rajgopal, 2011), political connections (Correia, 2014), and private-sector scrutiny, salient public events, and SEC personnel incentives (Holzman et al., 2024). Consequently, AAER-based datasets contain systematic label noise, as the absence of an AAER may reflect enforcement constraints rather than the absence of fraudulent or questionable reporting. This concern is well documented. Karpoff et al. (2017) show that research conclusions depend on the misconduct database employed, while Donelson et al. (2021) demonstrate that relying on a single enforcement regime materially changes empirical inferences and recommend broadening fraud measures beyond AAERs. More broadly, the objective of fraud detection is not merely to predict observed enforcement outcomes, but to identify firms exhibiting fraud-like characteristics that warrant closer scrutiny. Models capable of recovering this latent signal can help regulators prioritize investigative resources and provide investors and auditors with earlier warnings of potential financial reporting misconduct.

In this paper, we distinguish the latent economic object of interest from the observed enforcement label. Following Dechow et al. (2011), we define *fraud* as a material accounting misstatement that violates GAAP and is made with the intent of misleading investors, and denote this latent construct by F^*. This definition is a property of a firm's reporting conduct rather than of any regulatory outcome. By contrast, the data provide only an enforcement label, L, indicating firms that the SEC both detects and formally identifies through an Accounting and Auditing Enforcement Release (AAER). Accordingly, L is an incomplete and selectively observed proxy for F^*. Our objective is therefore to recover signals that are informative about the underlying construct beyond the subset revealed through AAERs.

The mismatch between the latent construct F^* and the observed label L gives rise to two fundamental challenges for fraud detection: model training and model evaluation. A model trained directly on L may learn patterns associated with the SEC's enforcement rather than the underlying phenomenon F^*. On the flip side of the same coin, conventional out-of-sample classification accuracy cannot determine whether a model has recovered information about F^* or merely reproduced historical enforcement patterns represented by L. Nor is this problem unique to SEC enforcement. Alternative labels based on private litigation are likewise selectively generated because securities class actions disproportionately target large firms experiencing salient stock-price declines (Donelson et al., 2021). More broadly, these challenges arise whenever observed labels are generated through an endogenous institutional revelation process rather than by direct observation of the underlying economic state.

Because the training problem motivates the design of our method, we discuss it first before turning to the evaluation problem. We approach the problem by adopting a simple principle: **future proves past**. The central idea is that future accounting outcomes occupy a qualitatively—not merely quantitatively—different informational position from contemporaneous accounting variables because they emerge only after earlier reporting choices have interacted with economic reality. If current reports are manipulated, contemporaneous accounting numbers may already embody those reporting distortions. Fu-

ture outcomes, by contrast, reflect what subsequently unfolds—operating performance, impairments, financing pressure, auditor scrutiny, and realized investment outcomes. Although noisy, these realizations may reveal tensions embedded in earlier financial reports that are not fully observable at the time they are issued. Because we apply this principle only within the training sample, however, model evaluation remains genuinely out-of-sample and free of look-ahead bias.

The intuition is fundamentally intertemporal. Fraud temporarily borrows from the future. Firms may inflate current performance through revenue inflation, expense understatement, or asset overstatement, but these accrual-based distortions cannot be sustained indefinitely and eventually reverse as economic reality catches up (Healy, 1985; Dechow et al., 1995). As the distortions unwind, subsequent accounting outcomes may exhibit characteristic patterns such as negative accruals, widening gaps between earnings and cash flow, deteriorating profitability, investment problems, and reductions in receivables or inventory (Dechow et al., 1996, 2011; Beneish, 1999; Kedia and Philippon, 2009).

Future outcomes, however, do not prove fraud. Many firms experience deteriorating performance or investment problems for ordinary economic reasons, while some fraudulent firms may avoid pronounced subsequent reversals. A procedure that mechanically relabels firms based on poor future performance would likely confuse fraud with financial distress, adverse shocks, or bad luck. Accordingly, our interpretation of the future-"proves"-past principle is deliberately modest. We use future outcomes only as an auxiliary source of information to make modest refinements to the supervisory signal during training, while AAER cases remain the anchor for confirmed enforcement actions. The objective is not to reclassify a new subset of firms as fraudulent ex ante, but rather to make the supervisory signal slightly more informative across the entire sample of firms. While each adjustment is modest, collectively they are intended to better align the observed training signal with F^* across the entire sample. The underlying intuition is that many machine learning methods learn from the structure of the entire training sample, so modest but broadly distributed improvements in alignment may be more valuable than large adjustments affecting only a small subset of firms.

While the training problem concerns how to recover a better supervisory signal from imperfect labels, the evaluation problem concerns how such models should be judged when those same labels are themselves incomplete. Simple assessment against L alone cannot establish whether a model has recovered information about F^* or merely reproduced historical enforcement patterns. We therefore evaluate the model using economically distinct validation tests outside the AAER label set. If the model successfully recovers information about F^*, firms predicted to be high risk should exhibit behaviors and outcomes independently associated with fraudulent reporting, even without formal enforcement actions. Thus, our framework combines complementary training and evaluation components: future accounting outcomes refine the supervisory signal during training, while economically distinct validation tests assess whether the recovered signal generalizes beyond observed enforcement actions.

Our sample period spans 1990–2019. We use windows ending by 1998 (with label information through 1999) as the initial training sample and reserve 2000–2019 for out-of-sample testing. A firm-year (i, t) receives a label of 1.0 if it is a confirmed AAER case. All remaining firm-years are initially assigned the neutral label of 0.5. We then refine these neutral labels for the *training* sample by incorporating accounting information from future years $\{t+1, t+2, t+3\}$, requiring that $t + 3 \leq 1999$. Specifically, we apply a logistic transformation to future accounting outcomes after orthogonalizing them with respect to firm size, producing the *Soft Label*.[1] Confirmed AAER cases retain a label of exactly 1.0, whereas non-AAER firm-years receive adjusted labels centered around the neutral value of 0.5.

Fraud detection presents a challenging learning problem because fraud signals are weak, noisy, high-dimensional, and exhibit substantial temporal structure. To extract these signals, we use a residual convolutional neural network (CNN) that performs firm-year fraud detection using the previous five years of 23 accounting variables, treating each firm-year as a 5×23 accounting matrix that preserves temporal structure and captures nonlinear interactions across variables and over time. We further incorporate cost-

[1] This "look forward" procedure is applied only within the training sample.

sensitive resampling and dropout regularization to improve statistical signal recovery under the highly imbalanced class distribution characteristic of fraud detection.

Our main approach combines the CNN architecture with the soft-label framework described above, which we refer to as the *Soft Label* CNN. As a benchmark, we estimate the same CNN architecture using only the conventional AAER labels, referred to as the *Hard Label* CNN. We evaluate all models using standard out-of-sample fraud detection metrics. In addition to the Hard Label CNN, we compare our approach against RUSBoost (Bao et al., 2020), a state-of-the-art machine-learning benchmark for fraud detection, and the Dechow et al. (2011) F-score, the standard accounting-ratio fraud predictor, re-estimated fully out of sample under the same expanding-window protocol. The Hard Label CNN provides a strong baseline and substantially improves detection performance relative to existing approaches in the literature, increasing detection rates by a factor of two or more. The Soft Label CNN, however, shifts the learning objective from reproducing observed SEC enforcement labels toward identifying latent fraud under selective enforcement. The validation tests that follow are designed to evaluate whether this shift in the learning objective improves the identification of latent fraud.

The first two validation tests use externally screened, relatively clean populations as negative benchmarks. Arthur Andersen client firm-years from 2002–2005 that emerged from the intense post-Enron forensic scrutiny without enforcement actions provide a post-scrutiny benchmark of accounting quality. Bank-loan holders identified through DealScan provide a second benchmark because they have passed creditor screening and remain subject to ongoing covenant monitoring.

Consistent with its objective of identifying latent fraud rather than reproducing enforcement outcomes, the Soft Label CNN systematically avoids flagging both groups as high risk. Within the large-firm (> \$750M) sample, post-scrutiny Arthur Andersen clients are 3.4 percentage points less likely to be flagged than observably similar peers (fractional-logit $z = -2.07$; randomization inference $p = 0.017$), making the Soft Label CNN the only model to exhibit a statistically significant tilt at conventional levels. Similarly, only 5.5% of bank-loan holders are flagged, compared with 15.0%

of non-borrowers. In contrast, all enforcement-trained alternatives reverse this pattern, flagging bank-loan holders more frequently than non-borrowers. The third validation test examines seasoned equity issuance. Firms with temporarily overstated performance may have incentives to issue equity before the overstatement reverses (Loughran and Ritter, 1995; Teoh et al., 1998; Baker and Wurgler, 2002). Soft-label top-decile firms complete seasoned equity offerings at nearly three times the base rate (26.8% over the next three years versus 9.1% for the bottom 90%, a 17.7 percentage point difference) and subsequently experience large negative post-issuance abnormal returns (-11.4% annualized FF25-adjusted returns over the $[t+1, t+5]$ window). In contrast, the Hard Label and RUSBoost show little association with equity issuance, and although the size-neutral F-score does flag future issuers, none of the three exhibits comparable post-issuance underperformance.

The fourth validation test examines whether predicted fraud probabilities identify future stock underperformance. A long–short portfolio that shorts the top decile of predicted fraud probability and holds the bottom 90% earns economically and statistically significant factor-adjusted alpha under the Soft Label CNN (7.5% under the Fama–French five-factor model, significant at the 1% level), whereas comparable portfolios formed using the enforcement-trained alternatives do not.

The four validation exercises rely on conceptually distinct sources of evidence—firm behavior, audit outcomes, creditor screening, and stock returns—yet all point in the same direction. Additional evidence comes from investor sentiment. The Soft Label CNN's flag rate uniquely co-moves positively with the Baker and Wurgler (2006) investor-sentiment index, concentrating in the hard-to-value industries where sentiment-driven fraud should surface (Wang et al., 2010). An ablation of the training target itself pins down where the information resides: the same network retrained on the forward-looking component alone—a pure deterioration predictor—detects essentially no enforced fraud and reproduces almost none of the validation footprint, so the hidden-fraud signature emerges only from the *combination* of the enforcement anchor and the forward tilt, not from either component in isolation. Our results remain robust across alternative soft-

label constructions, ranking rules, scaling choices, and training configurations (Appendix B). Taken together, the evidence suggests that the future-proves-past principle provides a more informative learning signal than observed enforcement labels alone, enabling machine learning models to better identify the underlying latent fraud construct, F^*.

A key feature of our design is that all validation tests are conducted entirely within the non-AAER sample ($F^* \setminus L$), allowing us to assess whether the model recovers economically meaningful information beyond the observed enforcement set. Our contribution is therefore not simply a more accurate fraud detector. Existing methods are designed to recover observed enforcement outcomes, and evaluated on that objective they perform well. Our results instead argue for a change in perspective: once enforcement labels are recognized as an incomplete proxy for the underlying fraud process, the learning objective itself becomes an object of study. The soft-label framework developed here represents one step in that direction, but more broadly this perspective motivates new approaches to learning and evaluating latent fraud.

The paper contributes to four strands of literature. It builds on the literature on hidden fraud and selective enforcement (Dyck et al., 2024) by treating the wedge $F^* \setminus L$ as the learning target. It extends the accounting literature on fraud and misstatement detection (Beneish, 1999; Dechow et al., 1995, 2011) and the growing machine-learning literature in accounting (Larcker and Zakolyukina, 2012; Brown et al., 2020; Bao et al., 2020) by introducing a deep learning framework that exploits the temporal structure of multi-year accounting data. Finally, it connects to the literature on the economic consequences of fraud and misreporting (Kedia and Philippon, 2009; Feroz et al., 1991; Loughran and Ritter, 1995; Teoh et al., 1998), since our validation tests examine whether model-identified firms exhibit the characteristics of hidden fraud firms before, or in the absence of, formal enforcement actions, and whether they subsequently exhibit the economic consequences of fraud and misreporting.

The remainder of this paper is organized as follows. Section 2 describes the data and sample construction. Section 3 develops the soft-label training target, the CNN implementation, and the training procedure, and introduces the benchmark models. Section 4

presents the out-of-sample detection results and documents the size bias of enforcement-trained models. Section 5 presents the four-test validation battery: the Arthur Andersen and bank-loan screening tests, equity issuance and post-issuance returns, and long-short portfolio alphas. Section 6 ablates the training target into its components and examines the investor-sentiment predictions. Section 7 concludes.

2 Data and Sample Construction

2.1 Defining Fraud

Corporate fraud in the broad sense spans a wide range of misconduct: insider trading, Ponzi schemes, benchmark manipulation, and the opportunistic timing of option grants, among others (Lie, 2005; Heron and Lie, 2007; Lie, 2025). Our object of interest is narrower: fraudulent *financial reporting*. Following Dechow et al. (2011), we define *fraud* as a material, economically significant accounting misstatement that (i) violates GAAP and (ii) is made with the intent of misleading investors. We refer to this latent construct as *latent fraud*. Fraud must be defined independently of SEC enforcement; otherwise, undiscovered fraud would be impossible by definition. Latent fraud is therefore not directly observable: some fraudulent misstatements are never detected, and some detected misstatements are never formally designated.[2]

Our empirical measure is the *enforcement label*: the part of latent fraud that the SEC both detects and designates through an Accounting and Auditing Enforcement Release (AAER). By construction, an AAER label is the *joint* outcome of three events—a misstatement occurs, the SEC detects it, and the SEC designates it in an enforcement release—so a firm-year carries the label only if all three hold (Dechow et al., 2011). The intent element of the definition enters the data only at the second and third stages:

[2]Following Dechow et al. (2011) and Dyck et al. (2024), we use the term *fraud* as shorthand for an intentional material misstatement. For the observed enforcement label defined below, the intent element is measured by the SEC's allegation, in the enforcement release, that the misstatement was intentional; consistent with the enforcement setting, the firms and managers involved typically neither admit nor deny guilt, and we make no claim of adjudicated guilt for any individual firm. Dechow et al. (2011) use the term *misstatement* for the same reason.

what we observe is the SEC's allegation that the misstatement was intentional, not intent itself, so enforcement is the measurement apparatus for latent fraud rather than part of its definition. The advantage of the enforcement label is its reliability: a firm-year that carries it is highly likely to be a genuine misstatement. Its limitation is censoring: Dyck et al. (2024) estimate that in normal times only about one-third of corporate fraud is ever detected, so the enforcement label is a selected, incomplete proxy for latent fraud.[3]

Fraud that occurred but escaped enforcement—the wedge between latent fraud and the enforcement label—is precisely the hidden fraud this paper sets out to detect (Figure 1 depicts this structure). The relation between the two is one of strict containment: every labeled firm-year is genuine fraud, but genuine fraud extends beyond the labeled firm-years. The training and evaluation problems of Section 1 are the two consequences of that containment being strict: training on the enforcement label risks learning the enforcement process rather than the fraud process, and accuracy measured against the enforcement label cannot certify detection of the fraud it leaves out.

One definitional caveat deserves emphasis at the outset. Latent fraud as defined above is narrow: it requires both a GAAP violation and intent, and neither element is observable outside the enforcement process. The forward-looking indicators underlying the soft label (Section 3.1) therefore cannot target latent fraud exactly; what they capture is a broader latent construct, aggressive accounting that subsequently reverses, which contains the hidden fraud we seek but may also contain within-GAAP earnings management that involves neither a violation nor demonstrable intent to mislead. This gap is not specific to our construction: because intent is unobservable, *any* accounting-based signal for hidden fraud is necessarily a signal for a superset of it. What can be assessed is whether the firms the signal ranks highest behave like fraud firms rather than like merely aggressive or distressed ones, and this is precisely the role of the validation exercises in Sections 5–6: opportunistic equity issuance ahead of reversals, avoidance of externally screened (relatively clean) populations, and negative risk-adjusted returns are patterns as-

[3]The incompleteness is measurable within the *detected* part of latent fraud itself: comparing public and private enforcement over 1998–2014, Donelson et al. (2021) find that roughly two-thirds (64%) of settled securities class actions alleging accounting fraud have no SEC enforcement counterpart. Even among frauds credibly detected by *some* channel, the AAER label designates only a minority.

sociated with misleading reporting rather than with aggressive-but-informative reporting or ordinary deterioration.

The remainder of this section describes how we operationalize the enforcement label from the AAER record; Section 3 then describes how the soft label uses forward-looking accounting outcomes to recover information about the fraud that lies beyond it.

2.2 Data Source

Our fraud sample is constructed from the Accounting and Auditing Enforcement Releases (AAERs) database compiled by Dechow et al. (2011). Since 1982, the SEC has issued AAERs during or at the conclusion of investigations against companies, auditors, or officers for alleged accounting and/or auditing misconduct. We use the version updated through December 31, 2021 (AAER No. 1 through No. 4,278), which covers 1,816 unique firm misstatement events. Among these, 1,087 events involve financial misstatements that affect at least one quarterly or annual financial statement. The annual file ("ann" sheet) records each fiscal year in which a misstatement occurred as a separate observation, yielding 2,195 firm-year observations across 942 distinct firms.

2.3 Sample Selection

Following Dechow et al. (2011), we apply the following filters to the annual file:

1. **Overstatement only.** We exclude firm-years in which earnings or revenues were understated, retaining only overstatement cases. This removes 56 firm-year observations (11 firms), leaving 2,139 firm-years (931 firms).

2. **Compustat identifier (GVKEY).** We require each firm to have a valid GVKEY for matching to Compustat. This removes 384 firm-years (187 firms), leaving 1,755 firm-years (744 firms).

3. **Compustat match with non-missing total assets.** We merge the remaining sample with Compustat annual data by GVKEY and fiscal year, and require total

assets (at) to be non-missing. This removes 136 firm-years (42 firms), leaving 1,619 firm-years (702 firms).

4. **Exclude financial firms.** We exclude banks, insurance companies, and other financial firms (SIC codes 6000–6999), as their financial statement structure differs substantially from non-financial firms and many accrual-related variables are not applicable. This removes 212 firm-years (93 firms), leaving 1,407 firm-years (609 firms).

After applying these four filters, we obtain 1,407 fraud firm-year observations from 609 distinct firms, covering fiscal years 1971 through 2019.

2.4 Sample Construction for Prediction

Our prediction model requires each fraud event to have five consecutive years of complete financial data ending in the misstatement year. We apply the following additional filters:

5. **Collapse to unique GVKEY firm-years.** The annual file records each fiscal year of each AAER as a separate observation, so nine records are duplicates at the firm-year level (two distinct AAERs covering the same fiscal year of the same firm), and several company records map to the same Compustat GVKEY. Collapsing to unique GVKEY firm-years leaves 1,398 fraud firm-years from 597 distinct GVKEYs.

6. **Analysis window restriction.** We restrict the sample to fiscal years 1990 through 2019, the analysis window used throughout the paper (Compustat coverage in earlier years is sparse, and the five-year rolling window requirement effectively bars pre-1985 events anyway). This removes 205 fraud firm-years (105 firms), leaving 1,193 fraud firm-years (492 firms).

7. **Five-year window requirement.** For each fraud event, we require at least five years of Compustat data up to and including the misstatement year. This removes 181 fraud events (85 firms) that lack sufficient history, leaving 1,012 events.

8. **Non-missing key variables.** We require 23 financial statement items to be non-missing across all five years of the window. The most frequently missing items are sale of stock ($sstk$), current assets (act), current liabilities (lct), and income taxes payable (txp). This removes 229 fraud events (102 firms).

9. **Consecutive fiscal years.** We require the five years in each window to be consecutive (i.e., no gaps in fiscal year). Once duplicate records are collapsed, this requirement drops no additional events.

2.5 Final Sample

After applying the Section 2.3 filters and the five-year window requirement, the AAER side of the panel contains **783 AAER fraud firm-years from 305 distinct firms** in the 1990–2019 analysis window. Each observation (i, t) uses the input window $\{t-4, \ldots, t\}$ (the five consecutive years of accounting data fed to the CNN) and is paired with a label window $\{t+1, t+2, t+3\}$ used to construct the soft label (Section 3.1). Note that the input window and the soft-label window are separated by only one year (input ends at t, label starts at $t+1$); we describe the additional safeguard that prevents this short label-window gap from leaking test-year accounting data into training in Section 3.4.3. The non-fraud side of the panel requires soft-label coverage: the label construction (Section 3.1) is defined on the CRSP NYSE/AMEX/NASDAQ common-equity universe (share codes 10/11; excludes ADRs, REITs, closed-end funds, ETFs, preferred securities, and OTC/pink-sheet stocks), and a five-year window enters the panel when any of its years carries a label (the most recent labeled year is used). Financial firms are in effect excluded by the 23-variable input requirement itself (banks and insurers lack several of the required working-capital items), so the resulting panel is more than 99% non-financial and non-utility at the terminal year. The full panel used for training and prediction comprises 84,089 firm-years (783 AAER fraud, 83,306 non-fraud); the out-of-sample test window 2000–2019 contains 54,115 of these firm-years (528 AAER firm-years; see Table 3). Table 2 reports the annual distribution; fraud rates peak in the early 2000s coinciding with the Enron/WorldCom era and decline steadily thereafter.

Our dataset exhibits severe class imbalance, a fundamental characteristic of fraud detection research. As shown in Table 2, fraud firm-years peak at 2.2% of the analysis sample (in 2001) and never exceed that share in any other year, and the pooled fraud rate is 0.93%. This extreme imbalance motivates the soft-label design that provides a continuous signal across non-AAER firm-years (Section 3.1) and the data-augmentation strategy described in the training procedure (Section 3.4).

3 The Soft Label and Model Design

3.1 Soft Label Construction

The standard approach in fraud detection trains models using binary labels: a firm-year is labeled 1 if it is subject to an SEC Accounting and Auditing Enforcement Release (AAER) and 0 otherwise, the *hard label* method. Section 2.1 established why this target is contaminated: Dyck et al. (2024) estimate that at least 10% of large publicly traded firms engage in fraud, while AAER records cover only about 1% of firms in our sample, so the "0" class mixes genuinely clean firms with undetected fraudulent ones, and the binary label treats them identically.

We propose a *soft label* method that operationalizes the future-proves-past principle of the Introduction: it preserves the AAER label on known AAER firms while replacing the uniform "0" target for non-AAER firms with a continuous measure of forward-looking accounting anomaly. Rather than using future stock returns, which conflate fraud-related information with a wide range of non-accounting shocks, we construct the soft label from forward-looking *accounting* variables that directly capture the correction patterns (accrual reversals, profitability deterioration, and write-downs) through which manipulation unwinds.

In the language of modern machine learning and econometrics, this construction simultaneously addresses three interrelated identification problems. **First, noisy-label correction**: the binary AAER label is contaminated in one direction (hidden fraud mislabeled as zero), and the soft label functions as an auxiliary supervisory signal

that partially de-noises the non-AAER class (Natarajan et al., 2013). **Second, weak supervision**: the four forward-looking indicators are not exact fraud labels but *noisy proxies* of the latent fraud state, in the spirit of distantly supervised and Snorkel-style labeling functions (Ratner et al., 2017). Each indicator individually is a weak signal of misreporting; their aggregation into an integer count c of binary flags converts four weak signals into a more reliable composite. **Third, latent-state inference under selective enforcement**: SEC enforcement is itself selective and size-biased, so the observed AAER label is a non-random sample of the latent fraud population (Dyck et al., 2024). The forward indicators are generated outside the SEC's selection process and thus partially purge the selection bias from the supervisory signal. The four-step construction below operationalizes these ideas jointly.

Step 1. Four forward-looking accounting indicators. For each firm-year (i, t), we construct four indicators $Y_{1,i,t}$ through $Y_{4,i,t}$ that capture forward-looking accounting reversal and deterioration patterns. We adopt a uniform sign convention: *lower values indicate worse (more fraud-like) outcomes*. The four indicators are:

- Y_1 *cash-flow-statement accruals*: $Y_{1,i,t} = -\,\text{Accruals}_{i,t}$, where $\text{Accruals}_{i,t} = (ib_t - oancf_t)/at_{t-1}$ is the gap between net income and operating cash flow scaled by lagged total assets. Large positive accruals (earnings well above cash flow) signal earnings management; the sign flip ensures large accruals map to low Y_1.

- Y_2 *accrual reversal*: $Y_{2,i,t} = -(\text{Accruals}_{i,t} - \text{Accruals}_{i,t-1})$, the year-on-year change in $\text{Accruals}_{i,t}$ defined above. A growing accruals gap maps to low Y_2.

- Y_3 *ROA deterioration*: $Y_{3,i,t} = \text{ROA}_{i,t} - \text{ROA}_{i,t-1}$, with $\text{ROA}_{i,t} = ib_t/at_{t-1}$. ROA falling year-on-year produces a negative Y_3, consistent with the convention.

- Y_4 *unexplained investment*: $Y_{4,i,t} = -|\widehat{\varepsilon}_{i,t}|$, where $\widehat{\varepsilon}_{i,t}$ is the residual from the within-(year, FF48 industry) cross-sectional OLS

$$\frac{capx_{i,t}}{at_{i,t-1}} = \alpha_{j,t} + \beta_{1,j,t}\, Q_{i,t} + \beta_{2,j,t}\, CF_{i,t} + \beta_{3,j,t}\, Size_{i,t} + \varepsilon_{i,t}, \qquad (1)$$

with Tobin's Q $Q_{i,t} = (|prcc_f_{i,t}| \cdot csho_{i,t} + lt_{i,t})/at_{i,t}$, cash flow $CF_{i,t} = oancf_{i,t}/at_{i,t-1}$, and size $Size_{i,t} = \log(at_{i,t})$; j indexes Fama–French 48 industries (financials and utilities excluded), and each (year, industry) cell with at least 20 non-missing observations is fit separately. Investment decisions poorly explained by Q, cash flow, and size produce large $|\widehat{\varepsilon}_{i,t}|$ and hence low Y_4.

Each $Y_{k,i,t}$ is winsorized at the 1st and 99th percentiles. Each indicator is then *residualized*: for every fiscal year we regress $Y_{k,i,t}$ cross-sectionally on a control vector containing the Altman Z-score, return on assets, leverage, log total assets, market-to-book, cash holdings, the trailing six-month stock return, and the trailing twelve-month return volatility, plus Fama–French 48 industry indicators, and replace $Y_{k,i,t}$ with the regression residual (the lower-is-worse orientation is preserved; we retain the notation $Y_{k,i,t}$ for the residualized values in the steps below). This regression stage purges the components of the raw indicators that are predictable from size, distress, valuation, liquidity, and recent stock-price behavior, so that the forward window aggregates *abnormal* reversals rather than mechanical correlates of firm condition; the within-(year, size-decile) ranking of Step 3 then adds a non-parametric size guarantee on top.

Step 2. Forward three-year aggregation. For each base year t and each indicator $k \in \{1, \ldots, 4\}$, we aggregate $Y_{k,i,s}$ across the strictly forward three-year window $\mathcal{T}_{i,t} \subseteq \{t+1, t+2, t+3\}$, requiring non-missing values in at least two of the three years. Indicators capturing accruals magnitude or year-on-year accounting reversals (Y_1, Y_2, Y_3) are aggregated by the within-window *minimum*: earnings-management episodes typically appear as a single charge-laden year, so the worst observed year is most informative, and a within-window mean would let a positive year cancel a severely negative one. The unexplained-investment residual (Y_4) is aggregated by the within-window *mean*: investment deviations from fundamentals tend to unwind gradually over multiple years, so sustained levels are more informative than single-year spikes. Formally,

$$\bar{Y}_{k,i,t} \;=\; \begin{cases} \min_{s \in \mathcal{T}_{i,t}} Y_{k,i,s}, & k \in \{1,2,3\}, \\ \frac{1}{|\mathcal{T}_{i,t}|} \sum_{s \in \mathcal{T}_{i,t}} Y_{k,i,s}, & k = 4. \end{cases} \tag{2}$$

Step 3. Size-decile-conditional ranking and binary flagging. A direct cross-sectional ranking of $\bar{Y}_{k,i,t}$ would mechanically tilt the soft label by firm size if any of the four aggregated indicators carries a size-related drift. To produce a label that is size-orthogonal *by construction*, we form size deciles within each fiscal year on $\log(at_{i,t})$ at the base year, and then rank each indicator within each (year, size-decile) cell:

1. For each base year t, partition firms into ten equal-sized bins by $\log(at_{i,t})$, indexed by $d(i,t) \in \{1, \ldots, 10\}$.

2. Within each (year, size-decile) cell with at least 30 observations, compute the within-cell empirical-CDF rank of $\bar{Y}_{k,i,t}$ for each k separately, with rank ascending in $\bar{Y}_k$ (so the lowest 10% by rank are also the 10% lowest values of $\bar{Y}_k$, which by the Step 1 sign convention are the most fraud-like on indicator k).

3. Define $F_{k,i,t} = \mathbf{1}\{\text{within-cell rank of } \bar{Y}_{k,i,t} \leq 0.10\}$, equal to one when firm i falls in the worst-10% sub-cell on indicator k relative to its size-decile peers in base year t, and zero otherwise.

By construction, the flag rate is exactly 10% per indicator within every (year, size-decile) cell, so any size-tilt in the underlying $\bar{Y}_k$ values is absorbed by the cell stratification rather than propagating to F_k. We then aggregate the four flags into a single integer count:

$$c_{i,t} \;=\; \sum_{k=1}^{4} F_{k,i,t} \;\in\; \{0, 1, 2, 3, 4\}, \tag{3}$$

where firm-years with any missing flag are dropped to ensure a clean count interpretation.

Step 4. Logistic transformation and AAER anchor. We map the count $c_{i,t}$ to a soft label via a centered logistic transform with scaling parameter $\beta = 0.05$, centered at the median count $c = 2$ so that $c_{i,t} = 2$ yields exactly 0.5:

$$\text{Soft Label}_{i,t} \;=\; 1 \;-\; \frac{1}{1 + \exp\left(-(2 - c_{i,t}) \cdot \beta\right)}. \tag{4}$$

Concretely, Soft Label $\in \{0.475, 0.488, 0.500, 0.512, 0.525\}$ for $c \in \{0, 1, 2, 3, 4\}$. AAER firm-years (those with $\text{AAER}_{i,t} = 1$ in the training data) are assigned a label of

exactly 1.0, overriding the logistic value. Non-AAER firm-years receive the count-based logistic value, which lies in a narrow band $[0.475, 0.525]$ around the neutral 0.5.

Two design choices deserve note. First, the binarization in Step 3 accepts some information loss, relative to a continuous within-cell z-score composite, in exchange for a clean size-orthogonality guarantee: each F_k has a flag rate of exactly 10% within every (year, size-decile) cell, so the count c cannot mechanically inherit any size tilt from the underlying $\bar{Y}_k$ distributions. Second, the small β keeps the soft label a gentle modifier of the neutral 0.5 target rather than a dominant signal that could overwhelm the AAER anchor at 1.0. We use the soft label as the training target; the hard label (binary AAER indicator) serves as the within-architecture benchmark throughout our analysis.

Methodological Position

Because hidden fraud is fundamentally unobservable, the contribution of the paper cannot rest primarily on classification accuracy or on the exact specification of the soft-label construction. In a latent-state environment with contaminated labels there is no verifiable record of actual fraud against which alternative constructions can be definitively compared; recovering latent accounting manipulation must therefore rely on imperfect proxy signals and weak supervision rather than exact classification. The informational role of the soft label derives not merely from injecting additional accounting variables but from the fact that the auxiliary signal is *forward-looking* rather than contemporaneous. This distinction is qualitative, not merely quantitative, and is the precise sense in which future proves past. Contemporaneous accounting variables are generated within the same reporting regime as any potential manipulation; if a firm is misreporting, those numbers *are* the misreported numbers. Future accounting outcomes occupy a fundamentally different informational position, because intertemporal accounting consistency and underlying economic constraints eventually exert pressure on the reporting process: revenue inflation, expense understatement, overstated asset values, and overly optimistic investment cannot be sustained indefinitely without future correction, reversal, or operational deterioration. Future realizations therefore act as partial *reality constraints* on the

latent accounting state behind earlier reports.

The empirical results below are not driven by a narrow specification choice. We conduct sensitivity analyses varying the indicator composition (substituting alternative correction-pattern indicators), the aggregation rule (MIN vs. MEAN), the within-cell ranking threshold (top 5%, 10%, 20%), the logistic scaling parameter β, the size-decile granularity, and the training configuration (architecture variants, fraud-replication multiplier, and certainty-weight bounds; the full ten-configuration hyperparameter grid and the pre-committed validation rule that selects the reported configuration are documented in Appendix B). The core qualitative findings remain robust across these alternative constructions; the contribution does not rest on a finely tuned specification but on the broader idea that future accounting reversals contain economically meaningful information about latent accounting problems that standard hard-label training procedures leave unexploited.

Figure 2 summarizes the resulting information timeline: for each training observation, future accounting information enters the training *label* only, never the model inputs, and the leakage-control rules described in Section 3.4.3 guarantee that no information dated at or after the test year influences any training label.

The central empirical burden of the paper therefore falls on the validation exercises of Sections 5–6, which assess whether the resulting signal behaves like hidden fraud rather than generic deterioration or distress; Section 6.1 subjects the target itself to the same discipline, ablating it into its enforcement-anchor and forward-tilt components and showing that each component alone fails exactly where the combination succeeds.

3.2 Dechow F-score Benchmark (Out-of-Sample Re-estimation)

Alongside the Hard Label CNN and RUSBoost (Bao et al., 2020), we include the Dechow et al. (2011) F-score (their Model 1) as a fourth benchmark. It is the standard accounting-ratio fraud predictor in the literature. Because it is a parsimonious logit on seven financial-statement ratios and uses no firm-size variable, it provides a transparent, size-neutral comparison for the machine-learning models. To place it on exactly the same

footing as the Soft Label, Hard Label, and RUSBoost models, we do *not* import Dechow et al.'s published coefficients; instead we *re-estimate* the model out of sample, under the identical expanding-window, retrain-per-year protocol used for the neural networks, on the identical firm-year sample and with the same binary AAER label.

Model 1 uses seven financial-statement variables (RSST accruals, the change in receivables, the change in inventory, the percentage of soft assets, the change in cash sales, the change in return on assets, and an actual-issuance indicator), each constructed from Compustat exactly as specified in Dechow et al. (2011), Table 3. For each test year $\tau \in \{2000, \ldots, 2019\}$ we estimate the logit by maximum likelihood on the expanding window of all real (non-replicated) firm-years with $1990 \leq t \leq \tau - 2$; the fitted coefficients are then applied to the held-out firm-years with $t = \tau$ to produce out-of-sample predicted probabilities. The six continuous variables are winsorized at their 1st and 99th percentiles, with the cut-points estimated on the training sample only. In all cross-model comparisons we rank firms within each year on the predicted probability (equivalently on the F-score, since the two are monotone within a year) and restrict the output to the identical test universe used by the other models; because Model 1's inputs are occasionally missing, the F-score is defined for 53,798 of the 54,115 test firm-years (99.4%) and is ranked within year on the non-missing set.

Before deploying the F-score as an out-of-sample benchmark, we verify that our variable construction reproduces Dechow et al. (2011): estimating their three nested logit models in sample on the strict-paper universe (overstatement AAERs with case number $\leq$ 2261; all Compustat firm-years 1979–2002) recovers coefficients that match the published Table 7 values in sign and magnitude across all three models and reproduces the paper's headline diagnostics, including an Enron (fiscal 2000) Model-1 F-score of 2.65 against the paper's 2.76. This confirms that the Model-1 inputs we feed into the out-of-sample re-estimation are built correctly.

3.3 Model Architecture

The choice of a Convolutional Neural Network is motivated by the requirements of the latent-fraud detection problem rather than by a general preference for deep learning. Our central empirical challenge is not simply class imbalance but *label contamination*: the observed AAER label is a selectively observed enforcement outcome, while many non-AAER firms may contain hidden fraud. Addressing this problem requires a framework that can ingest soft, probabilistic, or weak-supervision targets for non-AAER firms; standard hard-label classifiers, including RUSBoost (Bao et al., 2020), are designed around binary outcomes rather than the graded targets that are central to our approach.

A second consideration is that detecting hidden fraud is a low-signal, high-noise prediction problem. The relevant information is unlikely to reside in any single accounting ratio or isolated year; latent manipulation more plausibly appears through nonlinear interactions among accounting variables and through their evolution over multiple years.

The CNN provides such a framework. We represent each firm-year as a 5×23 matrix (five consecutive years $\times$ 23 Compustat variables; Table 1, Panels A–D), so that the network can learn local patterns across variables and over time without requiring all interactions to be specified ex ante (Figure 3 sketches the architecture). This flexibility is especially important in our setting because the goal is not merely to reproduce the observed SEC enforcement decisions, but to recover latent accounting-risk patterns that enforcement may miss. The CNN therefore serves as an implementation device that allows the soft-label idea to operate in a high-dimensional, nonlinear, and temporally structured environment. The architecture combines two convolutional blocks with a residual connection (LeCun et al., 2015; He et al., 2016) and a focal-loss objective adapted to the mixed hard/soft label structure; Figure 3 illustrates the complete architecture, and Appendix A reports the implementation details and the certainty-weighted focal-loss specification.

A deliberate feature of the architecture is the asymmetric treatment of the variable and time dimensions. The convolutional kernels span adjacent *variables* within a single fiscal year (kernels of height one in the year dimension), with weights shared across

the five years, while all *cross-year* linkages are formed by the position-specific fully connected layers that follow. For this adjacency to be economically meaningful, the 23 variables are arranged in financial-statement order (assets, then liabilities and equity, then income-statement items, then cash-flow and market items; Table 1, Panels A–D), each block in standard accounting sequence, so that columns adjacent in the input matrix are related items: cash beside receivables beside inventory, or sales beside cost of goods sold. The within-year kernel therefore spans coherent local groups of accounts rather than an arbitrary column ordering. The economic logic is that the two dimensions carry different kinds of structure. Within a fiscal year, financial-statement variables are bound by accounting identities and ratio relationships (accruals against operating cash flow, receivables against sales) that mean the same thing in every year of the window; sharing the within-year filters across years imposes exactly this stationarity and conserves parameters. Across years, by contrast, the input window is anchored at the prediction year: a misstatement sits at the end of the window, so the same accounting dynamic carries different information in year $t-1$ than in year $t-4$ as borrowed prosperity builds toward the terminal year. Position-specific dense weights preserve this calendar-position information. This is also why we do not convolve over the time dimension. A temporal convolution would impose time-translation invariance, forcing an accrual build-up four years before the prediction date to receive the same weight as one immediately preceding it, and thereby sacrificing exactly the calendar-position information that matters. With only five time steps, moreover, such a convolution is a constrained special case of the dense layer we already use, so it would add no expressive power in return. Keeping the temporal stage in the dense layers, with low capacity and heavy regularization, also limits overfitting risk given only 783 positive training events.

At the same time, the paper's contribution does not rest on the CNN architecture alone: the architecture is valuable because it enables the broader weak-supervision design, and the same design is in principle compatible with alternative function approximators suitable for tabular time-series inputs (e.g., transformer-style attention models, recurrent networks, or multi-layer perceptrons).

3.4 Training Procedure

3.4.1 Data Augmentation

To address class imbalance, we replicate every confirmed-fraud (AAER = 1) training window twenty times, so each fraud case enters training as twenty-one copies. Dropout regularization (Srivastava et al., 2014) ensures that each replicated copy is seen through a different effective subnetwork, so the replication functions as data augmentation rather than mere duplication of identical inputs.

3.4.2 Loss Function

During training, we combine observations with varying label certainty: hard labels ($y_i = 1$) for confirmed AAER fraud cases and soft labels ($y_i \in [0.475, 0.525]$) for non-AAER firm-years. To learn from this heterogeneous label structure, we use a Smooth Certainty-Weighted Focal Loss (Lin et al., 2017) that down-weights observations whose labels are uncertain (close to 0.5) and up-weights confident labels (close to 0 or 1):

$$\mathcal{L}_i = w(y_i) \cdot \alpha \, (1 - p_t)^{\gamma} \cdot \mathrm{BCE}(y_i, \hat{y}_i), \qquad w(y_i) = w_{\min} + (w_{\max} - w_{\min}) \cdot 2|y_i - 0.5|, \tag{5}$$

where $\hat{y}_i$ is the predicted probability, BCE is the binary cross-entropy loss, and $p_t = \exp(-\mathrm{BCE})$ is the predicted probability of the true class. The certainty weight $w(y_i)$ rises from $w_{\min}$ at $y_i = 0.5$ (maximum uncertainty) to $w_{\max}$ at $y_i \in \{0, 1\}$ (hard labels). Under our soft-label design, confirmed AAER cases receive the maximum weight, while non-AAER firm-years receive weights that increase monotonically with the magnitude of their forward-looking accounting-anomaly signal. Optimization uses the Adam optimizer with mini-batch gradient descent.

3.4.3 Out-of-Sample Scheme and Soft-Label Leakage Control

We use an expanding-window training scheme: for each test year $T \in \{2000, \ldots, 2019\}$, the model is trained on observations ending in $\{1990, \ldots, T-2\}$, with a two-year (24-

month) gap between training and test following Bao et al. (2020); we feed the model a five-year input window of the prior accounting data, and evaluate on observations whose input window ends at T. The aligned out-of-sample test sample contains 54,115 firm-years and 528 AAER firm-years (Table 3). For each test year, we follow Bao et al. (2020) and classify the top 1% of firms by predicted probability as flagged. The 1% cutoff mirrors their rationale: because typically fewer than 1% of firms in a year are AAER fraud cases, flagging the top 1% by predicted probability targets a fraction comparable to the base rate of enforced fraud, rather than relying on a classifier-specific probability threshold that would depend on the model and the training sample.

The two-year input-side gap is necessary but not sufficient when the training target is the soft label, because the soft label of a training observation with input window ending at t is computed from accounting variables in $\{t+1, t+2, t+3\}$ (Section 3.1, Step 2). For training samples with $t \in \{T-2, T-3\}$, this label window overlaps the test year T (e.g., $t = T-3$ yields label window $\{T-2, T-1, T\}$), so the unmasked soft label would carry information about year-T accounting outcomes that the model would not have at prediction time. We close this gap by replacing the soft label of any non-AAER training sample with $t \in \{T-2, T-3\}$ by the neutral value 0.5 during the training of test year T, removing all directional signal from those observations. AAER firm-years are unaffected because their label is the binary AAER indicator and does not depend on future accounting data. Training samples with $t \leq T-4$ have label window $\subseteq \{T-3, T-2, T-1\}$ and are used with their actual soft label.

4 Fraud Detection Performance

We organize the out-of-sample evidence into two distinct sets that play different roles in the paper. The first, the *detection* results of this section (Table 3 and Figures 4–5), compares model rankings directly against the SEC labels and measures each model's ability to recover known AAER cases. These tests serve as a necessary check that the soft-label target does not sacrifice SEC-labeled fraud capture. The second, the *validation*

results of Sections 5 and 6 (Tables 4–10), compares model rankings against external signals that lie outside the SEC label set entirely: post-scrutiny Arthur Andersen status, bank-loan screening, future equity issuance behavior, post-issuance returns, long-short portfolio alpha, and investor-sentiment exposure. Because no data source records which firms committed the fraud that enforcement missed, the validation set carries the central identification burden of the paper: detection establishes that the soft label preserves SEC-labeled fraud capture, while validation establishes that it captures the hidden fraud the SEC misses. Throughout, four methods are compared on the same test universe: our Soft Label CNN, the identically specified Hard Label CNN, RUSBoost (Bao et al., 2020), and the out-of-sample re-estimated Dechow et al. (2011) F-score (Section 3.2).

4.1 Sample Characteristics

Table 2 presents the annual distribution of unique five-year sample windows from 1990 to 2019. The Step 1 pool comprises 135,897 firm-years from Compustat with a complete, consecutive five-year window of the 23 accounting variables listed in Table 1; the soft-label-coverage requirement (which embeds the CRSP common-equity screen of the label construction) reduces this to the 84,089-firm-year analysis sample used for training and prediction. The 783 AAER misstatement years (already filtered to non-financial in Table 2) are retained throughout. Annual coverage at Step 1 ranges from 3,969 windows (2019) to 5,301 (1999); the unconditional AAER fraud rate peaks in 2001 during the Enron/WorldCom era and falls sharply after 2010. This extreme class imbalance motivates both the soft-label design described in Section 3.1 and the fraud-case replication strategy described in Section 3.4.

For our out-of-sample evaluation, we train on data through fiscal year $T-2$ and test on T for each test year $T \in \{2000, \ldots, 2019\}$. The aligned test sample is the subset of the 84,089-firm-year prediction panel that corresponds to test years 2000–2019; it contains 54,115 firm-years and 528 AAER firm-years, and is the basis for the detection results in Table 3. The cross-sectional and portfolio analyses (Tables 4–10) draw on subsets of this aligned sample as detailed in each section.

4.2 Detection Results

Table 3 reports out-of-sample fraud detection results for 2000–2019 on the aligned sample of 54,115 firm-years (528 AAER fraud firm-years). We compare the Soft Label CNN (our proposed method), the Hard Label CNN (the same architecture trained with binary AAER labels), RUSBoost (Bao et al., 2020), and the out-of-sample Dechow et al. (2011) F-score. For each test year we classify the top 1% of firms by predicted fraud probability as flagged. Each model's score is its within-year ranking averaged across the 30 training runs, the run-noise-robust ensemble aggregate used for all ranking-based results in the paper. The table reports both recall (the detection rate, TP/(TP+FN)) and precision (TP/flagged); because the within-year top-1% rule flags the same fraction of firms for every method, the two measures move together.

One scope caveat applies to every model compared here: all four see only accounting numbers, so the relevant benchmark is the accounting-detectable part of latent fraud. Enforcement cases triggered by non-accounting conduct—textual disclosure, governance breaches, market manipulation—or by manipulation too subtle to surface in the financial statements lie outside the reach of any accounting-based model, however flexible, and place an upper bound on the detection rates reported below.

On the full aligned sample (Panel B), the yearly-mean detection rate is 17.3% for the Soft Label CNN, 17.9% for the Hard Label CNN, 6.1% for RUSBoost, and 7.7% for the F-score; pooled detection rates are 18.0%, 18.6%, 8.1%, and 4.4% respectively. The two neural-network models are broadly comparable (Hard slightly ahead, consistent with the hard label being trained directly on the evaluation target), and both detect known AAER cases at two to four times the rate of RUSBoost and the F-score. Restricting the evaluation to large firms with total assets above $750 million (Panel A)—the symmetric Dyck et al. (2024) threshold—sharpens this pattern: Soft and Hard detect roughly 19–21% of large-firm AAER cases per year while RUSBoost and the F-score fall below 4%, a first indication that the benchmarks' full-sample detection interacts strongly with firm size, a pattern we examine directly in the next subsection.

One caveat concerns the label itself: the detected counts thin out in the later sample

years partly because SEC accounting enforcement itself contracted sharply after the mid-2000s (Donelson et al., 2021), and recent misstatement years are further truncated by the enforcement lag. The label, not the models, loses power late in the sample.

These figures answer the primary detection question: the soft-label target gives up little SEC-labeled fraud capture relative to its hard-label twin, and both neural models recover known AAER cases at substantially higher rates than the machine-learning and statistical benchmarks. But they leave open the more interesting empirical question of *which* firms each model flags. Because the two neural models differ only in their training target (soft vs. hard labels) and are applied to the same test sample, any systematic differences in the rankings they produce are informative about how the soft label changes model behavior. We investigate this in Section 4.3 and in the validation tests of Sections 5–6.

4.3 Size-Controlled Discrimination and the Size Bias of Enforcement-Trained Models

With clean labels, the area under the curve (AUC) has a simple ranking interpretation: it is the probability that a randomly chosen fraud firm-year is ranked above a randomly chosen non-fraud firm-year, and its partial version (pAUC) applies the same comparison within the strictest part of the ranking, where a fraud screen must operate. Two features of our setting complicate that interpretation. First, the measure treats every non-AAER firm-year as fraud-free, so a model that successfully surfaces hidden fraud is penalized for doing so: each hidden-fraud firm-year it ranks highly is scored as a false alarm. Second, SEC enforcement is heavily tilted toward larger firms, so a model trained on hard AAER labels inherits that size tilt and ranks large firms higher than a size-orthogonalized model would.

Figure 4 plots the yearly-averaged normalized pAUC for all four models as the false-positive-rate (FPR) cutoff is swept from 0% to 50%, across three test windows (2000–2008, 2000–2014, 2000–2019) and two benchmark pools: the full non-AAER pool and the pool restricted to firms with total assets of $750 million or more, which holds firm

size roughly fixed. In the full pool the AAER-trained Hard Label leads, the Soft Label is competitive, and RUSBoost catches up only at lenient (high-FPR) cutoffs. Under the size-restricted benchmark, however, RUSBoost's discrimination weakens sharply (over 2000–2019, its pAUC at the 50% FPR cutoff falls from 0.56 to 0.41, and at the strict 5% cutoff from 0.12 to 0.06), whereas the Soft Label is essentially unchanged (0.47 to 0.52) and stays on par with the Hard Label. This contrast traces directly to the training target: RUSBoost, estimated against the enforcement label, inherits its size profile, so part of its full-pool discrimination reflects firm *size*; the size-neutral Soft Label discriminates on information that survives size control.

Figure 5 shows the size bias directly. Among each model's within-year top-decile flagged firm-years, the median flagged firm has total assets of $143 million under the Soft Label, close to (indeed slightly below) the $347 million population median, versus $468 million under the F-score, $809 million under the Hard Label, and $1.59 billion under RUSBoost, with roughly 70% of RUSBoost's flags above the $750 million cutoff. The enforcement-trained models concentrate their flags on large firms, reproducing the selection pattern of the enforcement process they are trained on; this is the empirical counterpart of the training problem described in the Introduction, and it motivates the size-controlled designs used in the validation tests below.

5 Validating the Hidden-Fraud Interpretation

The validation tests of this section are designed as a battery, not a list, because any single test can be passed by a model that captures something other than fraud. Each test therefore draws on a different source of information about hidden fraud—ex-post regulatory and audit scrutiny (Section 5.1), ex-ante creditor screening (Section 5.2), the firm's own financing behavior (Section 5.3), and market prices (Section 5.4)—and we require the conclusion "the soft-label model captures hidden fraud" to hold simultaneously across all four; no single alternative signal produces all four results jointly. Each subsection below states the economic mechanism behind its test before presenting the results. As

the results will show, the Soft Label is the only one of the four models that aligns on every test. This is not a deficiency of the alternatives: the F-score and RUSBoost were built (and our Hard Label CNN was trained) to agree with the enforcement label, an objective on which they perform exactly as designed; where a benchmark departs from the hidden-fraud predictions below, we trace the departure to its design—its training target, its size profile, or its inputs—rather than leave it unexplained.

The joint requirement also rules out the most plausible innocent reading of the flags: that the model simply identifies firms in financial distress or in ordinary business decline. That reading makes predictions of its own, and on two of the four tests they run opposite to ours: distressed firms face depressed share prices and limited access to equity markets, so a model that ranked them highest would flag firms that issue equity *less* often than the average firm, not more; and a portfolio formed on distress should earn little abnormal return once size, value, and profitability are held fixed, because those factors already absorb most of the compensation for distress risk. Section 6.1 confronts this alternative directly rather than by implication: it retrains the identical network on the forward deterioration signal alone, with the enforcement anchor removed, and shows that the resulting model detects essentially no enforced fraud and reproduces almost none of the footprint documented below.

5.1 Validation Against Post-Scrutiny Arthur Andersen Clients

Our first validation test asks whether each model's flags tilt toward or away from a population subjected to unusually intense *ex-post regulatory and audit scrutiny*. When Arthur Andersen collapsed in 2002, all of its corporate clients had to be re-audited by a successor auditor (predominantly the surviving Big 4), and a coordinated regulatory response and intense media spotlight focused on whether AA's prior audit work had concealed accounting irregularities. A non-AAER firm that survived this scrutiny is not certified fraud-free, but as a group the post-scrutiny AA cohort carries a lower residual hidden-fraud rate than the unscrutinized non-AA pool—a *relatively* clean anchor, and that weaker premise is all the test requires. The prediction is directional: a model that

captures hidden fraud should tilt its high-risk flags *away* from this population, while a model that has merely inherited the size tilt of SEC enforcement will flag it *more*, because AA's former clients lean toward larger firms.

Table 4 implements the test on this post-scrutiny anchor. Both panels pool the Arthur Andersen client firm-years of 2002–2005 (identified by the 2001 audit signature, following Dyck et al., 2024), exclude AAER cases, and restrict to large firms with total assets above $750 million—the canonical Dyck–Morse–Zingales sample—so that the anchor comparison is like-for-like on size and cannot be driven by the mechanical size tilt documented in Figure 5. The test is non-circular: the anchor is defined by the audit signature, not by any fraud label the models trained on.

Panel A is descriptive: within the large-firm pool, the 431 post-scrutiny AA firm-years are spread across each model's ranking roughly like the non-AA population, with RUSBoost placing the most in its top decile (48, or 11.1%) and the two neural models the fewest (34 each, or 7.9%). Panel B is the formal test: a fractional-logit regression of the *flag probability* (the share of 30 independently trained runs that place the firm-year in the within-year top decile of the > $750M pool) on the AA-client indicator, with firm controls, year fixed effects, size-decile×industry fixed effects, and firm-clustered standard errors; coefficients are log-odds. The Soft Label is the only model whose coefficient is significant at conventional levels: post-scrutiny AA clients' log-odds of being flagged are 0.43 lower than observably similar large firms' (-0.427, $z = -2.07$), an average marginal effect of -3.4 percentage points against a 10% baseline flag rate. The estimate is corroborated by randomization inference: across 300 firm-level reassignments of the AA label within industry×size strata, refitting the model on each draw, the observed coefficient is more negative than 98.3% of the placebo draws (one-sided $p = 0.017$). The Hard Label coefficient is smaller and insignificant (-0.352, $z = -1.57$), RUSBoost's is positive and insignificant ($+0.159$, $z = 0.57$), and the F-score's is insignificant (-0.317, $z = -0.89$).

We emphasize what this test does and does not show: no model actively over-flags the anchor once size is held fixed, and the stacked equality tests reject only RUSBoost's

coefficient against the Soft Label's at the 10% level. The value of the test is directional and joint: the Soft Label is the only model that tilts its flags away at conventional (5%) significance from a cohort of firm-years subjected to unusually intense regulatory and audit scrutiny, which is what a genuine fraud signal (as opposed to an enforcement-salience signal) should do.

5.2 Bank-Loan Screening

The second validation test replaces ex-post regulatory scrutiny with *ex-ante creditor screening*, an entirely different source of information about hidden fraud. Banks accumulate private information through credit due diligence and continuing covenant monitoring—non-public reviews of customer–supplier relationships, internal control documents, and ongoing financial-statement reviews performed for covenant compliance—and they have strong financial incentives to identify accounting risk, because their own capital is at stake. The two mechanisms are complementary: one is ex-post and regulatory, the other ex-ante and contractual, so together they provide a much stronger negative benchmark than either could alone. If the Soft Label CNN captures accounting-quality patterns that such monitors screen on, firms with outstanding syndicated bank loans (which have passed bank due diligence and accept ongoing covenant monitoring) should be flagged less often than firms without such loans.

Table 5 implements this test. We build a firm-year indicator *has_loan* using DealScan linked to Compustat via the Chava–Roberts link, and restrict to non-AAER firms over 2000–2019.

Panel A reports the raw flag rates. Under the Soft Label, 5.5% of loan-holding firm-years fall in the within-year top decile of predicted fraud probability, versus 15.0% of firm-years without an active loan. Under every enforcement-trained benchmark the pattern reverses: loan holders are flagged *more* often than non-holders (11.7% vs. 8.2% for the Hard Label, 12.6% vs. 7.2% for RUSBoost, and 12.4% vs. 7.3% for the F-score), reflecting the concentration of bank borrowers among the larger firms where those models' flags concentrate (Figure 5). Panel B confirms the Soft Label result

in a multivariate logistic regression (log-odds coefficients) controlling for the Graham et al. (2008) bank-loan determinants (leverage, profitability, market-to-book, tangibility, the standardized Altman Z-score, and cash-flow volatility) and for cash holdings, which belong in this specification because committed credit lines substitute for cash (Sufi, 2009), plus year and size-decile×industry fixed effects: the *has_loan* coefficient is negative and significant only under the Soft Label (-0.268, $z = -3.77$; an average marginal effect of about -2.0 percentage points), and statistically zero under the Hard Label (-0.096, $z = -1.43$), RUSBoost (-0.100, $z = -1.35$), and the F-score (-0.074, $z = -1.14$). The contrast with Panel A is informative: the enforcement-trained benchmarks' higher raw flag rates for borrowers vanish once size deciles are held fixed (their Panel A patterns reflect size composition rather than borrower risk), whereas the Soft Label avoids loan holders in both the raw rates and the controlled specification.

In sum, firms that have passed bank due diligence, a classic ex-ante screening mechanism, are systematically avoided by the Soft Label's high-risk flags. Combined with the AA-client analysis, this constitutes consistent cross-sectional evidence: firms that outside professional monitors independently screen (bank-monitored or post-scrutiny AA) are flagged less by the soft-label model, while the enforcement-trained alternatives flag them at similar or higher rates. Which approach appears preferable therefore depends on the evaluation criterion: models built to agree with the enforcement label rank highest against that label, while externally screened benchmarks favor the model trained to look beyond it.

5.3 Future Equity Issuance

The two screening tests above are negative benchmarks: populations a genuine fraud signal should avoid. The third validation test provides the battery's *positive* behavioral signature, drawn from the firm's own financing behavior. The economic mechanism is well established in the seasoned-equity-offering literature: firms whose accounting overstates true performance face a closing window before the overstatement reverses, and they rationally exploit the temporary overvaluation by issuing equity (Loughran and

Ritter, 1995; Teoh et al., 1998; Baker and Wurgler, 2002); once the accounting unravels through subsequent write-downs, earnings disappointments, or operational reality, the recently issued equity underperforms benchmark returns. This issue-and-underperform footprint is observable directly from market data and mechanically independent of SEC enforcement: a hidden fraudster who issues equity in t+1 and underperforms thereafter generates exactly the same observable footprint as an SEC-prosecuted fraudster would. If the Soft Label CNN identifies firms with unsustainable accounting performance, its top-decile firms should therefore be disproportionately likely to issue equity in the period immediately following the prediction, and the equity they issue should subsequently underperform; a detector trained on SEC labels alone, by contrast, has no direct mechanism through which to learn this issuance-timing signal.

5.3.1 Issuance Rates and Cross-Sectional Logit Evidence

Table 6 reports both raw issuance rates (Panel A) and logit estimates (Panel B) for the strictly future $[t+1, t+3]$ window, where issuance is a completed SDC-recorded seasoned equity offering (IPOs and convertibles excluded). In Panel A, Soft Label top-decile firms complete an SEO at a 26.8% rate over the next three years versus 9.1% in the bottom 90%, nearly three times the base rate ($t = 28.63$). By contrast, Hard Label top-decile firms issue at the unconditional rate (11.0% vs. 10.9%, $t = 0.27$), RUSBoost top-decile firms issue *less* often than the bottom 90% (9.1% vs. 11.1%, $t = -4.71$), as expected given the large, mature firms its flags concentrate on (Figure 5), and the F-score shows a modest positive gap (12.1% vs. 10.4%, $t = 3.62$), consistent with its size-neutral accounting-ratio design carrying part of the issuance signal.

Panel B reports a logistic regression (log-odds coefficients) with Issue$(t+1..t+3)$ as the dependent variable and a top-decile *High Risk* indicator as the regressor of interest, controlling for lagged equity issuance, leverage, ROA, Tobin's Q, momentum, book-to-market, and the seasoned-equity-offering determinants of DeAngelo et al. (2010) (cash holdings, firm age, R&D, idiosyncratic volatility, the Frank and Goyal (2003) financing deficit, and a dividend-payer dummy), plus year and size-decile×industry fixed

effects. The High Risk coefficient is +0.400 ($z = 5.77$) under the Soft Label and +0.248 ($z = 3.16$) under the Hard Label—a roughly 49% versus 28% increase in the odds of a completed SEO—statistically zero ($z = 0.71$) under RUSBoost, and +0.417 ($z = 5.64$) under the F-score. The lagged-issuance control (which loads at $\beta \approx 0.9$, $z \approx 16$) absorbs persistent-issuer behavior, so the coefficients are interpreted as the *newly emerging* component of issuance risk: the Soft Label's coefficient is roughly 1.6 times the Hard Label's, RUSBoost shows no incremental issuance association once size, industry, and the issuance determinants are absorbed, and the F-score's log-odds coefficient is comparable to the Soft Label's. The proportional-odds scale compresses the Soft Label's advantage relative to Panel A because its flags sit in cells with high baseline issuance propensity; the level gap in Panel A remains far larger for the Soft Label (17.7 percentage points versus 1.7 for the F-score), while the two models' average marginal effects are comparable (untabulated: +4.3 versus +4.5 percentage points) only because they are evaluated against very different baselines.

5.3.2 Post-Issuance Returns

Table 7 reports post-issuance Fama–French 25-adjusted stock returns for firms that are both in the top decile of each model's predicted probability *and* issue equity. Under the Soft Label, top-decile equity issuers deliver economically large negative annualized FF25-adjusted returns over the post-issuance $[t+1, t+5]$ window: -11.4% per year on the value-weighted FF25 benchmark and -11.1% on the equal-weighted benchmark, both significant at the 1% level. The corresponding averages are -3.5% for Hard Label issuers and -3.5% for F-score issuers (both about half the -6.6% earned by the average issuer in the full sample), and $+0.3\%$ (indistinguishable from zero) for RUSBoost issuers. Restricting to "Soft-only" firms (top-decile under Soft Label but not under Hard Label) sharpens the result to -12.8% (VW) and -12.5% (EW) over the same window.

The combined picture from Tables 6–7 is consistent with the soft-label design: the top-ranked Soft Label firms behave like firms with unsustainable accounting performance who opportunistically time equity issuance and then subsequently underperform the

FF25 benchmark by economically large amounts, even though only a small fraction of them are ever subject to SEC enforcement. Among the benchmarks, it is the size-neutral F-score, not the enforcement-trained learners, that shows a partial version of this footprint (a positive issuance gap and -3.5% post-issuance returns), which reinforces the interpretation: the issue-and-underperform signature is carried by accounting-based fraud signals, and the Soft Label's contribution is to sharpen that signal severalfold rather than to introduce a pattern absent from the accounting literature.

5.4 Economic Significance: Long-Short Portfolio Alphas

The final test of the battery moves from sub-populations to market prices: it is the only test that integrates over the full cross-section rather than focusing on a particular group of firms, the only one that uses stock returns rather than accounting variables, and the only one that requires the model's predictions to be informative *after* adjusting for priced risk. The economic mechanism connecting hidden fraud to portfolio returns is gradual price adjustment: as fraud-related information slowly diffuses into prices through earnings disappointments, write-downs, and management changes, the firm's stock underperforms benchmark returns. If the Soft Label CNN successfully identifies firms with unsustainable accounting performance (both detected and undetected), then a portfolio that shorts the top-decile high-risk firms and holds the remaining 90% should produce positive returns as those firms' borrowed prosperity reverses. The Fama–French factor adjustment is essential to the test: it strips out returns attributable to standard risk premia (market, size, value, momentum, profitability, investment), so that any remaining alpha must come from fraud-related information that the factor model does not already capture.

Table 8 reports Fama–French factor regressions of monthly long-short portfolio returns for the four scoring methods: Soft Label (Panel A), Hard Label (Panel B), RUSBoost (Panel C), and the F-score (Panel D). Each year firms are independently ranked by each method and the top 10% are shorted while the remaining 90% are held long, equal-weighted, with a three-year overlapping holding window (signal years 2000–

2019) under the Fama–French timing convention: a signal from the fiscal year ending in calendar year T enters the portfolio at the end of June of $T+1$, so every constituent's annual report is public before it is traded on. Constituent monthly returns are delisting-adjusted and winsorized at the 1st/99th percentiles before portfolio formation. AAER firms are excluded from the portfolios. Alphas are annualized and reported in percent; t-statistics use Newey–West standard errors with 12 lags, reflecting the serial correlation induced by the overlapping holding period.

The Soft Label portfolio earns an annualized three-factor alpha of 13.0% ($t = 4.86$) and a four-factor alpha of 12.1% ($t = 4.78$). The five-factor specification absorbs a large positive loading on RMW of 0.82 ($t = 10.05$), consistent with the strategy shorting lower-profitability firms, and the alpha declines to 7.5% but remains significant at the 1% level ($t = 3.09$). In contrast, none of the alternatives earns significant alpha in any specification: the Hard Label's three-factor alpha is −0.2% ($t = -0.33$), RUSBoost's is −0.3% ($t = -0.24$), and the F-score's is +0.4% ($t = 0.44$). Across all specifications, the Soft Label alpha exceeds every alternative by roughly 7 to 13 percentage points per year.

This pattern, comparable SEC-label detection rates but sharply different portfolio alphas, completes the four-test battery and directly supports the soft-label design. Hard Label and Soft Label both capture known AAER cases at similar rates (Table 3), but only Soft Label's predictions carry risk-adjusted return information, consistent with the soft label capturing accounting anomaly patterns that translate into future underperformance irrespective of whether the SEC subsequently prosecutes.

6 Mechanisms and Additional Analyses

6.1 Ablating the Training Target: Anchor, Tilt, and Their Combination

The validation evidence admits one remaining alternative interpretation, which we state in its strongest form. The soft label is constructed from future accounting outcomes, so perhaps the model has learned nothing about fraud at all: it may simply predict *which*

firms are about to deteriorate, and frequent equity issuance, poor subsequent returns, and avoidance by screened populations might all be generic traits of firms headed downhill rather than signatures of concealed misreporting. If that reading were correct, the paper's central claim (that the soft-label model recovers hidden fraud) would reduce to a deterioration forecast.

We evaluate this interpretation by constructing the model it implies. Table 9 decomposes the soft-label target into its two components and evaluates each in isolation. The *de-anchored* CNN is the paper's own network (identical architecture, inputs, hyperparameter configuration, expanding windows, and 30-run ensemble) retrained on a target from which all enforcement information has been removed: AAER firm-years receive the same accounting count label as everyone else, the fraud-replication step is dropped, and the label-neutralization rule is applied to every window. What remains is exactly the forward tilt, so this network is the strongest feasible implementation of a pure deterioration predictor within our framework. Alongside it we evaluate a transparent linear version (an expanding-window forecast of the three-year-ahead ROA decline from current characteristics, estimated only on cohorts whose outcome window is complete) and, at the opposite vertex, the Hard Label CNN, which trains on the enforcement anchor alone.

The decomposition delivers a two-sided verdict. On the detection side, both tilt-only models are essentially blind to enforced fraud: at the top-1% cutoff the de-anchored CNN recovers only 1.3% of AAER firm-years, essentially the recall of a random ranking, against 17.9% for the Hard Label and 17.3% for the Soft Label (an untabulated linear ROA-decline forecast likewise recovers essentially none, 0.6%). The asymmetry is mechanical rather than incidental: manipulated reports make the manipulating firm's current accounting look *healthier*, so a signal built from "current statements predict decline" points away from fraudulent firms almost by construction. On the validation side, the tilt alone produces only a faint shadow of the Soft Label's footprint. The de-anchored CNN's excess SEO rate—how much more often its flagged firms complete a seasoned equity offering over the next three years than the rest of the pool—is only

+3.0 percentage points ($t = 6.22$), against +17.7 points ($t = 28.63$) for the Soft Label. Its flagged issuers' subsequent underperformance is likewise marginal (untabulated: $t = -1.84$, versus $t = -4.64$ for the Soft Label). The same asymmetry extends to the two screened-population tests: the de-anchored CNN's excess flag rate is essentially zero for post-scrutiny Arthur Andersen clients (+0.2 pp, against −2.6 pp for the Soft Label) and mild for bank borrowers (−1.5 pp against −9.5 pp), and in untabulated size-controlled logit regressions both of its coefficients are indistinguishable from zero (AA-client +0.063, $z = 0.21$; *has_loan* −0.011, $z = -0.17$), whereas the Soft Label's are significant (−0.427, $z = -2.07$; −0.268, $z = -3.77$; Tables 4 and 5).

Panel C completes the picture and provides the constructive half of the argument: the de-anchored CNN's top-decile flags overlap the Soft Label's by only 30.3%. In untabulated rank correlations, the Soft Label ranking nonetheless correlates substantially with the de-anchored CNN (Spearman +0.57), showing the forward tilt is genuinely present inside the soft-label model, while its correlation with the Hard Label ranking is just +0.10. Reading the three vertices together: the anchor alone detects enforced fraud but carries no hidden-fraud footprint; the tilt alone carries a faint footprint but detects nothing; their combination detects enforced fraud *and* produces the full footprint: an excess SEO rate of +17.7 percentage points where the components separately deliver +0.1 and +3.0. The hidden-fraud signature is therefore not the sum of its parts but an interaction: the anchor teaches the network what enforced fraud looks like in the accounting matrix, and the tilt redirects that representation toward firms outside the enforcement set whose accounting subsequently unravels. This is the paper's claim, that the training target is a first-order modeling choice, demonstrated within its own architecture, and it rules out the deterioration-forecast reading of the validation evidence: a model that only predicted decline would sit at the de-anchored vertex, and the de-anchored vertex fails every test that the Soft Label passes.

6.2 Investor Sentiment and the Predicted-Fraud Rate

Our final validation exercise moves from firm-level outcomes to the time series and cross-industry structure of the model's flags, asking whether they respond to an economic determinant of fraud that lies entirely outside any fraud label: investor sentiment. Theory and evidence connect the two: Wang et al. (2010) show that fraud propensity rises with investor beliefs about industry prospects, and Baker and Wurgler (2006) show that sentiment moves the prices of hard-to-value, hard-to-arbitrage firms the most. A model that genuinely tracks fraud incidence should therefore flag more firms in high-sentiment periods, and the effect should concentrate in sentiment-sensitive industries.

Figure 6 provides the aggregate, time-series view. Each model's annual flag rate (under its fixed in-sample top-decile cutoff) is plotted against the macro-orthogonalized Baker and Wurgler (2006) sentiment index.[4] The Soft Label flag rate co-moves with sentiment ($r = +0.72$); the realized SEC AAER count is also positively associated, consistent with genuine accounting-fraud incidence rising in high-sentiment periods; the Hard Label and F-score aggregates show no such association, while the RUSBoost aggregate also co-moves ($r = +0.50$)—a first sign that a single aggregate series cannot separate the models. With twenty annual observations, however, aggregate correlations are suggestive at best: they have limited power to discriminate among models, and a single time series offers no way to hold industry composition or business conditions fixed.

The theory's distinctive prediction is in any case cross-sectional: if sentiment drives fraud, flag rates should rise most in the industries whose firms are hardest to value (Baker and Wurgler, 2006). Table 10 therefore moves the test into an industry-year panel, where identification no longer relies on the aggregate series at all. Industry-year flag rates are regressed on the interaction of sentiment with each industry's sentiment-sensitivity expo-

[4] Baker and Wurgler (2006) orthogonalize each of the six underlying sentiment proxies to a set of macroeconomic conditions before forming the index: growth in industrial production; real growth in durable, nondurable, and services consumption; growth in employment; and an NBER recession indicator. The orthogonalized index therefore measures investor sentiment beyond contemporaneous business-cycle fundamentals, so the co-movement documented here is not a mechanical reflection of macroeconomic conditions.

sure; industry and year fixed effects absorb all main effects—the year effects subsume, in particular, any aggregate co-movement of the kind Figure 6 displays—and industry-year fundamentals enter as controls. This is the exposure-times-aggregate-shock design of Rajan and Zingales (1998). The exposure is built the same way in both column blocks (the industry share of firms in the market-wide top decile of a composite score, the overpriced tail in which Stambaugh et al. (2015) show sentiment effects concentrate), and the blocks differ only in the score: the left block uses the published SYY composite mispricing score itself, while the right block applies SYY's average-of-rankings aggregation to the five valuation-uncertainty characteristics of Hribar and McInnis (2012) (the two exposures correlate at 0.71 across industries). Every specification includes an exposure-specific linear time trend, $E_i \times (t - 2000)$, which rules out a spurious co-trending of hard-to-value industries' flag rates with the sentiment index. It also includes the industry business-conditions controls of the fraud-waves literature (Wang et al., 2010) (industry asset growth, sales growth, ROA, and the prior year's industry stock return), so that the interaction captures sentiment beyond real industry booms. The Soft Label is the only model that responds. A one-standard-deviation increase in sentiment raises the flag rate of a one-standard-deviation-more-exposed industry by 0.92 percentage points under the published-score exposure ($\delta = +0.921$, $t = 4.54$) and by 1.36 percentage points under the characteristics-composite exposure ($\delta = +1.362$, $t = 7.17$); the Hard Label and RUSBoost interactions are insignificant in both blocks, and the F-score's is insignificant under the characteristics exposure and marginally *negative* under the published-score exposure (-0.224, $t = -1.65$). Without the trend control the Soft Label estimates are larger still (untabulated: $+1.11$, $t = 3.74$ and $+1.74$, $t = 7.05$). RUSBoost shows a marginal positive interaction under the characteristics exposure only before the trend control is added (untabulated: $+0.28$, $t = 1.89$); the estimate is not robust to the trend control, to alternative exposure constructions, or to excluding the pharmaceuticals industry (see the table note), so we do not interpret it. Because sentiment is external to any fraud label and the exposures combine published ingredients with a published aggregation method, this is out-of-sample evidence: the soft-label model uniquely tracks the Baker and Wurgler

(2006) prediction (sentiment-driven fraud surfaces most in the industries whose firms are hardest to value), consistent with the industry-beliefs fraud channel of Wang et al. (2010).

7 Conclusion

This paper addresses a fundamental challenge in fraud detection: enforcement records reveal only a small fraction of actual fraud—Dyck et al. (2024) estimate that at least 10% of large publicly traded firms engage in fraud and that only about one-third of it is ever detected—so AAER-based training data are contaminated at scale by *hidden fraud*. Identifying these hidden fraudsters is the central goal of the paper, which we frame as a problem of recovering latent accounting manipulation under selectively observed enforcement labels. We replace the uniform zero target for unprosecuted firms with a soft label built from forward-looking accounting outcomes: a weakly supervised device that uses future accounting realizations as partial evidence on the latent state behind earlier reports.

Out of sample over 2000–2019, the soft-label model gives up essentially nothing in SEC-labeled fraud capture relative to its identically specified hard-label twin, while much of the benchmarks' apparent full-sample performance turns out to reflect firm size rather than fraud. The validation evidence, in contrast, separates the models sharply and points uniformly in one direction. Soft Label top-decile firms issue equity at nearly three times the base rate and subsequently underperform by economically large margins. The two externally screened populations—post-scrutiny Arthur Andersen clients and bank-monitored borrowers—are systematically avoided by the Soft Label's flags, a tilt no enforcement-trained alternative shows. A long-short portfolio on its ranking earns significant Fama–French alpha where every alternative earns none. Retraining the same network on the forward tilt alone removes both the detection ability and the footprint, so the hidden-fraud signature emerges only from the combination of enforcement anchor and forward tilt. And the flag rate co-moves with investor sentiment precisely in hard-

to-value industries, an economic determinant of fraud external to any label.

Because firms that committed undetected fraud are never revealed as such, no single test could establish that the soft-label model identifies them. The tests above are designed to align only if the model captures genuine hidden fraud: top-decile firms display the issue-and-underperform fraud footprint, two externally screened populations (post-scrutiny AA and bank-monitored) are avoided rather than over-flagged, the long-short portfolio earns positive Fama–French alpha, and the flags respond to sentiment where theory says fraud should. Taken together, this mutually supporting evidence (behavioral, ex-post regulatory, ex-ante creditor, stock-return, and macro-financial) suggests that the soft-label model's predictions reflect hidden-fraud capture rather than statistical artifacts.

Beyond the specific construction, the paper's broader message is a diagnosis. Once enforcement labels are recognized as an incomplete and selectively revealed proxy for fraud, agreement with those labels and capture of the underlying fraud process become separate objectives, and the choice of training target becomes a first-order modeling decision rather than a default. The soft label is one way—surely not the only way—to internalize the second objective; what our evidence establishes is that the two objectives genuinely diverge, that the divergence is empirically first-order, and that future accounting outcomes are informative enough to exploit it. We are therefore confident about the problem and deliberately modest about the solution: future accounting outcomes prove the past only weakly, and the discipline lies in using them accordingly.

References

Baker, Malcolm and Jeffrey Wurgler (2002) "Market Timing and Capital Structure," *Journal of Finance*, 57 (1), 1–32.

——— (2006) "Investor sentiment and the cross-section of stock returns," *The Journal of Finance*, 61 (4), 1645–1680.

Bao, Yang, Bin Ke, Bin Li, Y Julia Yu, and Jie Zhang (2020) "Detecting accounting fraud in publicly traded US firms using a machine learning approach," *Journal of Accounting Research*, 58 (1), 199–235.

Beneish, Messod D (1999) "The detection of earnings manipulation," *Financial Analysts Journal*, 55 (5), 24–36.

Brown, Nerissa C., Richard M. Crowley, and William B. Elliott (2020) "What Are You Saying? Using Topic to Detect Financial Misreporting," *Journal of Accounting Research*, 58 (1), 237–291.

Correia, Maria M. (2014) "Political Connections and SEC Enforcement," *Journal of Accounting and Economics*, 57 (2–3), 241–262.

DeAngelo, Harry, Linda DeAngelo, and René M. Stulz (2010) "Seasoned equity offerings, market timing, and the corporate lifecycle," *Journal of Financial Economics*, 95 (3), 275–295.

Dechow, Patricia M, Weili Ge, Chad R Larson, and Richard G Sloan (2011) "Predicting material accounting misstatements," *Contemporary Accounting Research*, 28 (1), 17–82.

Dechow, Patricia M, Richard G Sloan, and Amy P Sweeney (1995) "Detecting earnings management," *The Accounting Review*, 70 (2), 193–225.

——— (1996) "Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC," *Contemporary Accounting Research*, 13 (1), 1–36.

Donelson, Dain C., Antonis Kartapanis, John McInnis, and Christopher G. Yust (2021) "Measuring Accounting Fraud and Irregularities Using Public and Private Enforcement," *The Accounting Review*, 96 (6), 183–213.

Dyck, Alexander, Adair Morse, and Luigi Zingales (2024) "How pervasive is corporate fraud?" *Review of Accounting Studies*, 29 (1), 736–769.

Feroz, Ehsan H., Kyungjoo Park, and Victor S. Pastena (1991) "The Financial and Market Effects of the SEC's Accounting and Auditing Enforcement Releases," *Journal of Accounting Research*, 29, 107–142.

Frank, Murray Z. and Vidhan K. Goyal (2003) "Testing the pecking order theory of capital structure," *Journal of Financial Economics*, 67 (2), 217–248.

Graham, John R., Si Li, and Jiaping Qiu (2008) "Corporate misreporting and bank loan contracting," *Journal of Financial Economics*, 89 (1), 44–61.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016) "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Healy, Paul M (1985) "The effect of bonus schemes on accounting decisions," *Journal of Accounting and Economics*, 7 (1-3), 85–107.

Heron, Randall A and Erik Lie (2007) "Does backdating explain the stock price pattern around executive stock option grants?" *Journal of Financial Economics*, 83 (2), 271–295.

Holzman, Eric R., Nathan T. Marshall, and Brent A. Schmidt (2024) "When Are Firms on the Hot Seat? An Analysis of SEC Investigation Preferences," *Journal of Accounting and Economics*, 77 (1), 101610.

Hribar, Paul and John McInnis (2012) "Investor sentiment and analysts' earnings forecast errors," *Management Science*, 58 (2), 293–307.

Karpoff, Jonathan M., Allison Koester, D. Scott Lee, and Gerald S. Martin (2017) "Proxies and Databases in Financial Misconduct Research," *The Accounting Review*, 92 (6), 129–163.

Kedia, Simi and Thomas Philippon (2009) "The Economics of Fraudulent Accounting," *Review of Financial Studies*, 22 (6), 2169–2199.

Kedia, Simi and Shiva Rajgopal (2011) "Do the SEC's Enforcement Preferences Affect Corporate Misconduct?" *Journal of Accounting and Economics*, 51 (3), 259–278.

Kirkos, Efstathios, Charalambos Spathis, and Yannis Manolopoulos (2007) "Data mining techniques for the detection of fraudulent financial statements," *Expert Systems with Applications*, 32 (4), 995–1003.

Larcker, David F. and Anastasia A. Zakolyukina (2012) "Detecting Deceptive Discussions in Conference Calls," *Journal of Accounting Research*, 50 (2), 495–540.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015) "Deep learning," *Nature*, 521 (7553), 436–444.

Lie, Erik (2005) "On the timing of CEO stock option awards," *Management Science*, 51 (5), 802–812.

——— (2025) *Catching Cheats: Everyday Forensics to Unmask Business Fraud*: Berrett-Koehler Publishers.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár (2017) "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

Loughran, Tim and Jay R Ritter (1995) "The new issues puzzle," *Journal of Finance*, 50 (1), 23–51.

Natarajan, Nagarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari (2013) "Learning with Noisy Labels," in *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 1196–1204.

Rajan, Raghuram G. and Luigi Zingales (1998) "Financial dependence and growth," *American Economic Review*, 88 (3), 559–586.

Ratner, Alexander, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré (2017) "Snorkel: Rapid Training Data Creation with Weak Supervision," *Proceedings of the VLDB Endowment*, 11 (3), 269–282.

Ravisankar, Pediredla, V Ravi, G Raghava Rao, and Indranil Bose (2011) "Detection of financial statement fraud and feature selection using data mining techniques," *Decision Support Systems*, 50 (2), 491–500.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014) "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, 15 (1), 1929–1958.

Stambaugh, Robert F., Jianfeng Yu, and Yu Yuan (2015) "Arbitrage asymmetry and the idiosyncratic volatility puzzle," *The Journal of Finance*, 70 (5), 1903–1948.

Stambaugh, Robert F. and Yu Yuan (2017) "Mispricing factors," *The Review of Financial Studies*, 30 (4), 1270–1315.

Sufi, Amir (2009) "Bank lines of credit in corporate finance: An empirical analysis," *Review of Financial Studies*, 22 (3), 1057–1088.

Teoh, Siew Hong, Ivo Welch, and Tak Jun Wong (1998) "Earnings management and the long-run market performance of initial public offerings," *Journal of Finance*, 53 (6), 1935–1974.

Wang, Tracy Yue, Andrew Winton, and Xiaoyun Yu (2010) "Corporate fraud and business conditions: Evidence from IPOs," *The Journal of Finance*, 65 (6), 2255–2292.

Figure 1: The Iceberg Problem: Two Distortions from Enforcement-Based Labels

Schematic of the paper's setting. SEC enforcement (AAER) reveals only the tip of corporate fraud: our 2000–2019 evaluation sample contains 528 AAER firm-years, while the Dyck et al. (2024) pervasiveness estimates imply that only about one-third of fraud is ever detected, so at least twice as many fraudulent firm-years sit hidden below the waterline among the ≈53,600 unprosecuted observations. This structure creates two distinct distortions. The *training problem*: a model trained on hard AAER labels learns the properties of the visible tip—what enforcement selects, most notably large firms—rather than of fraud itself. The *evaluation problem*: any benchmark that treats non-AAER firm-years as verified negatives charges a model a "false positive" precisely when it surfaces hidden fraud. The paper's soft-label training target addresses the first distortion; its validation designs (post-scrutiny anchors, forward outcomes) address the second.

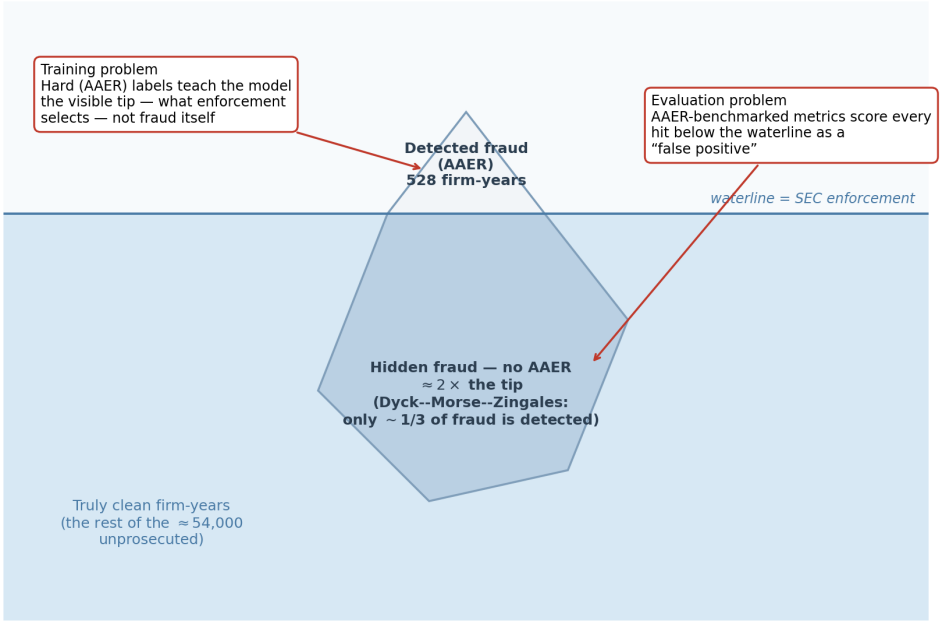


Figure 2: Soft-Label Construction Timeline: Where Future Information Can and Cannot Flow

Left: the information timeline for one test year τ. A training observation (i, t) feeds the model five years of inputs $[t-4, t]$ (blue); its soft label is built from the strictly forward window $\{t+1, t+2, t+3\}$ (orange), so future information enters the *training label only*, never the model inputs. (When the terminal-year label is unavailable—roughly a fifth of labeled non-AAER training windows, typically near data gaps or delisting—the most recent label within the window is used; most such fallbacks still draw on at least one post-t year, and in the small remainder (~5% of windows) the label is older, though all non-AAER labels stay within the narrow $[0.475, 0.525]$ band.) Training uses the expanding window $1990 \le t \le \tau-2$ (two-year gap), and the newest admitted cohorts ($t \in \{\tau-3, \tau-2\}$), whose label windows would touch the test year, have their non-AAER labels replaced by exactly 0.5 (gray), so no information dated τ or later influences any training label. The test prediction at τ uses inputs $[\tau-4, \tau]$ only. *Right*: the resulting label values. Non-AAER firm-years occupy a deliberately narrow band around 0.5 (from 0.475 for unflagged firm-years—71% of firm-years—to 0.525 for firm-years flagged on all four accounting indicators), while AAER firm-years keep the anchor label 1.0: a modest adjustment to the zero target, not a relabeling. *Bottom*: the training schedule across the full out-of-sample period, one row per test year $\tau = 2000, \ldots, 2019$: the model for τ trains on window-end cohorts $1990 \le t \le \tau-2$ (blue), with the two newest cohorts' non-AAER labels neutralized to 0.5 (gray), the gap year $\tau-1$ excluded, and predictions made at τ (red); the training window expands as τ advances.

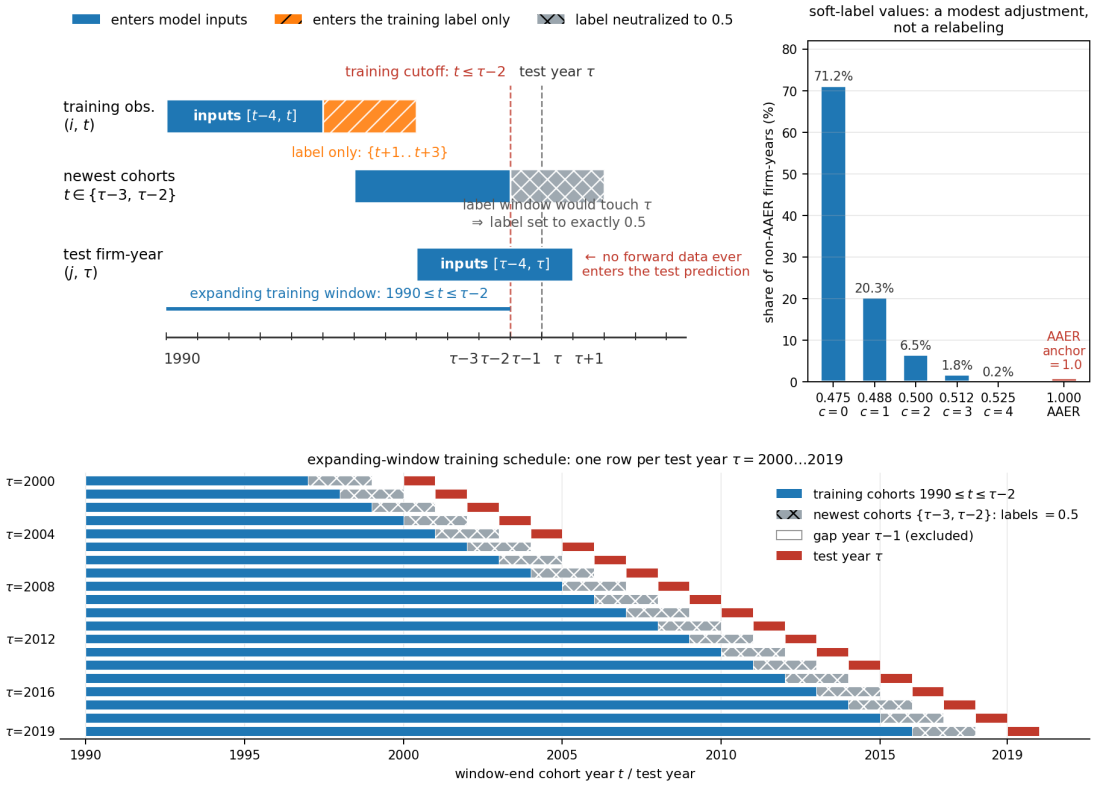


Figure 3: CNN-Based Fraud Detection Model Architecture with Residual Connections

The input is the 5×23 firm-year accounting matrix (five consecutive years of the 23 Compustat variables). Two convolutional blocks extract hierarchical features; the first includes a residual skip connection that facilitates gradient flow. The feature maps are flattened and passed through three fully connected layers to a sigmoid fraud probability. The certainty-weighted focal loss (inset) assigns higher weight to labels near 0 or 1 and lower weight to labels near 0.5, so the model learns jointly from hard AAER labels (anchored at 1.0) and the soft labels of non-AAER firm-years (a narrow band around 0.5).

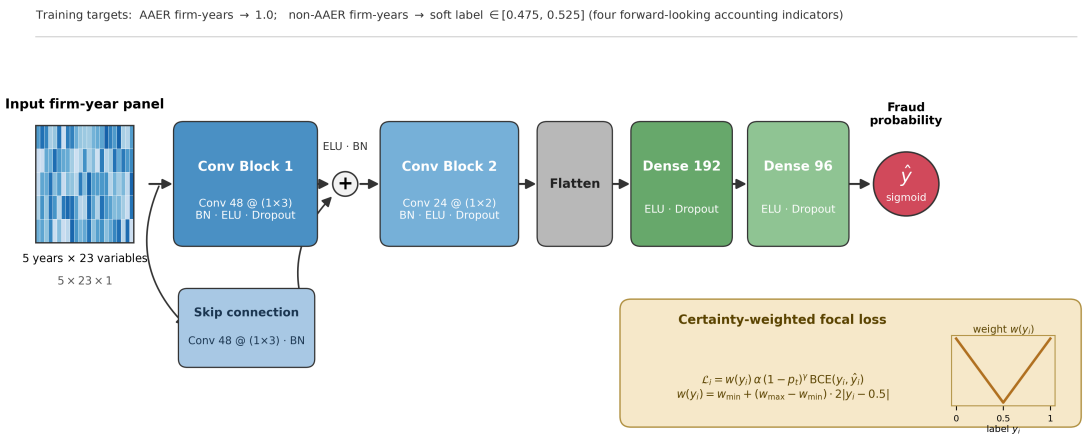


Figure 4: Detection Discrimination Dynamics: Normalized Partial AUC vs. FPR Cutoff (Within-Year, Size-Controlled)

Normalized partial AUC (pAUC) for separating AAER from non-AAER firm-years as a function of the false-positive-rate (FPR) cutoff, computed *within each fiscal year* and averaged across years. Rows: out-of-sample windows 2000–2008, 2000–2014, 2000–2019. *Left column*: non-AAER pool restricted to firms with total assets ≥ $750M (AAER firms retained), which holds firm size roughly fixed. *Right column*: full non-AAER pool. In the full pool the AAER-trained Hard Label leads, the Soft Label is competitive, and RUSBoost catches up only at lenient (high-FPR) cutoffs. Under size restriction RUSBoost's discrimination weakens sharply (2000–2019: pAUC at FPR=50% falls from 0.56 to 0.41, and at FPR=5% from 0.12 to 0.06), whereas the Soft Label is essentially unchanged (0.47→0.52) and stays on par with the Hard Label. Because pAUC depends only on the within-year ranking of predicted probabilities, the curves are identical whether computed from raw scores or from the within-year percentile ranks that underlie the paper's top-percentile flagging rule.

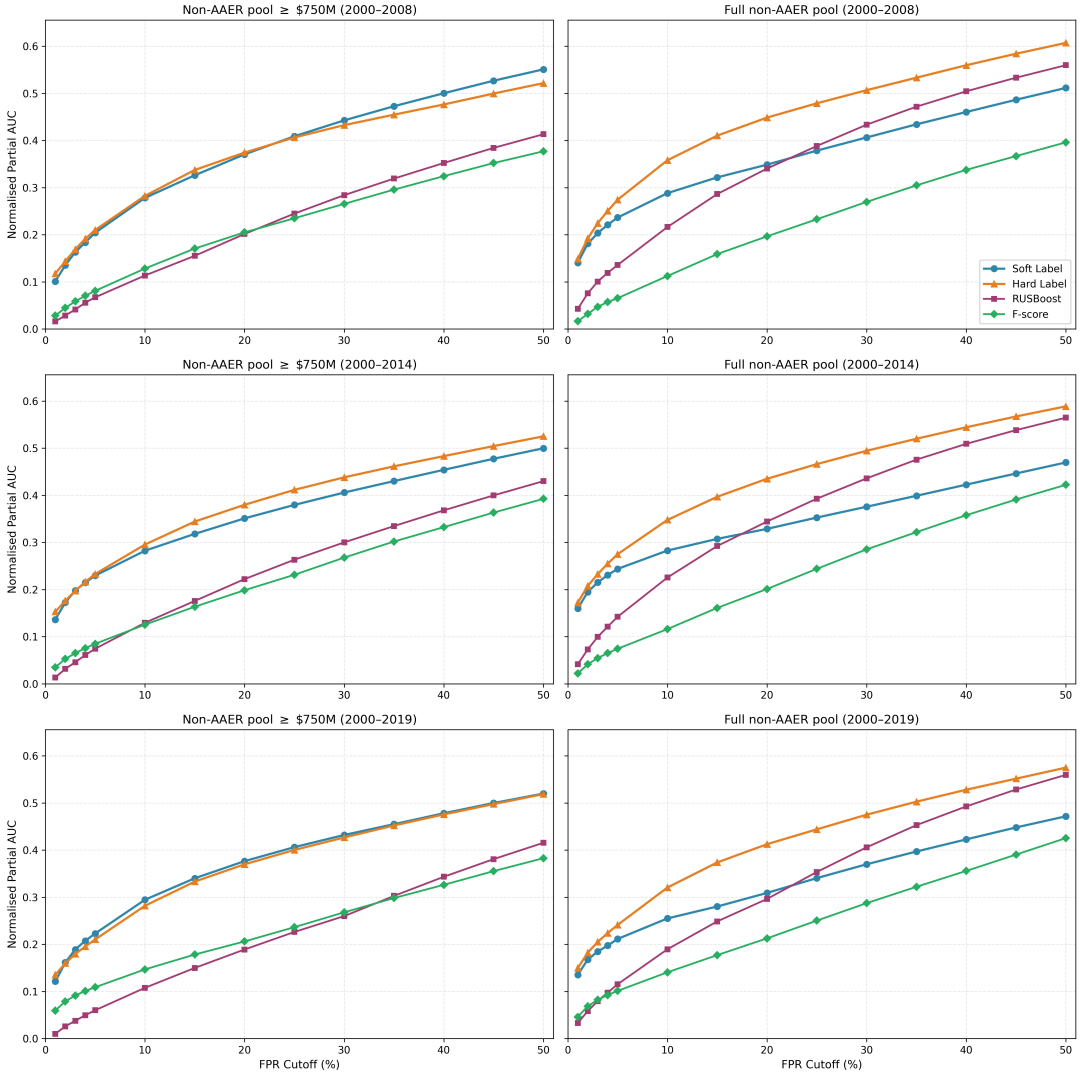


Figure 5: Firm-Size Distribution of Flagged Firms

Cumulative distribution of total assets among each model's within-year top-decile flagged firm-years (non-AAER, 2000–2019), against the distribution over all firm-years (dashed gray). The median flagged firm has total assets of $143M under the Soft Label, $468M under the F-score, $809M under the Hard Label, and $1.59B under RUSBoost, against $347M for the population; roughly 70% of RUSBoost's flags lie above the $750M Dyck–Morse–Zingales cutoff (dotted line). The enforcement-trained models concentrate their flags on large firms—the size bias inherited from the enforcement process, which motivates the size-controlled designs used throughout the paper—whereas the Soft Label's flags track (indeed slightly undershoot) the population size distribution.

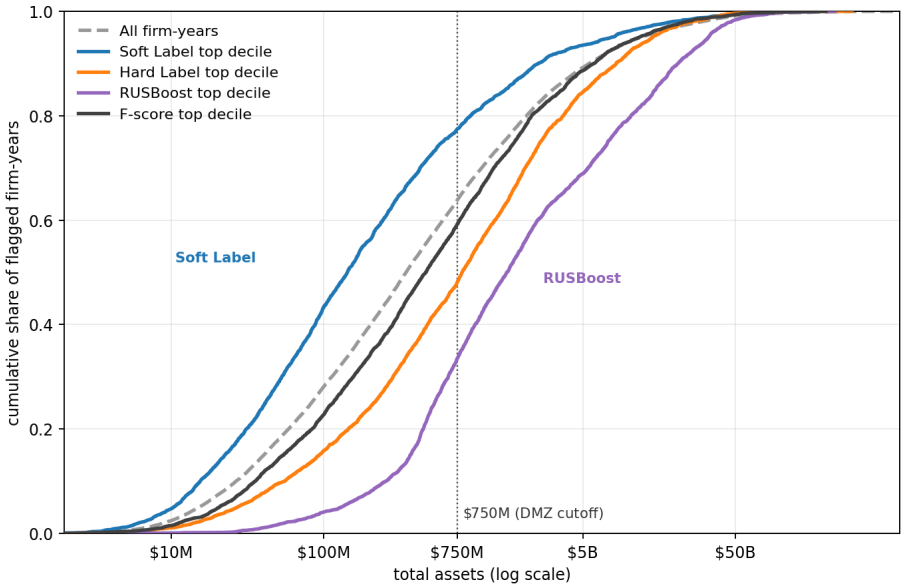


Figure 6: Baker–Wurgler Investor Sentiment and Aggregate Predicted Fraud Probability across Models

Annual aggregate fraud incidence versus the Baker and Wurgler (2006) investor-sentiment index orthogonalized to macroeconomic conditions (SENT_ORTH; dashed gray, right axis), 2000–2019. Panels A–D plot the annual fraction of non-AAER firms flagged by each model (left axis); Panel E plots the SEC AAER firm-year count as an external fraud-incidence reference. Firms are flagged against each model's *fixed in-sample* cutoff—the 90th percentile of its 2000-fold training-set scores, averaged over the 30 training runs—applied to every test year, so the threshold is calibrated out of sample; for the deterministic F-score the percentile is taken on Dechow's F-scale, whose raw probability scale drifts out of sample. Panel titles report the raw Pearson correlation with SENT_ORTH. The Soft-label flag rate co-moves with sentiment (r = +0.72), as does RUSBoost's (r = +0.50); the Hard-label and F-score aggregates do not. The AAER count is also positively associated, though its post-2014 decline partly reflects enforcement-lag truncation (the AAER data end 2021); the model flag rates are unaffected.

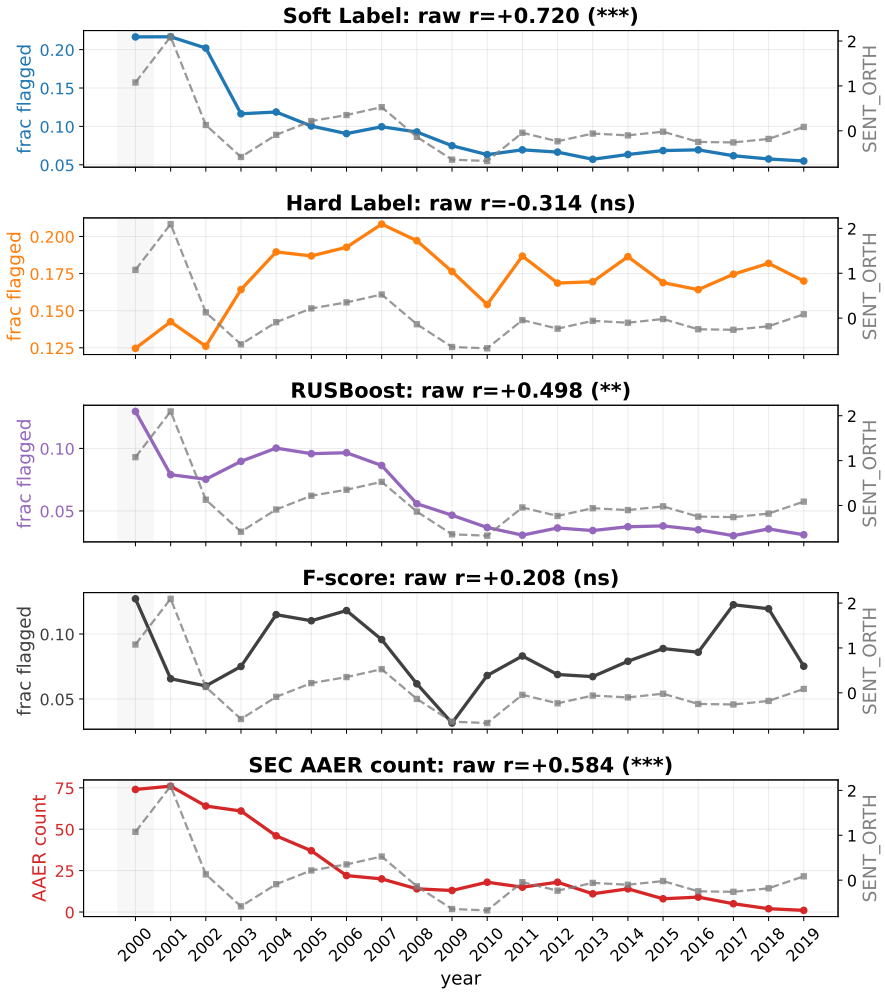


Table 1: Compustat Items and Variable Definitions

Panels A–D list the 23 raw Compustat (Annual) items used as model inputs; each firm-year requires non-missing values for all 23 items in all five rolling-window years (see Table 2, Panel B). Panel E defines the constructed variables and indicators used in the empirical tables. Ratios use Compustat Annual; returns from CRSP. AAER firms excluded from all analyses unless noted.

Variable	Definition
Panel A: Compustat Inputs — Assets (6 items)	
`che`	Cash and short-term investments
`rect`	Receivables, total
`invt`	Inventories, total
`act`	Current assets, total
`ppegt`	Property, plant, and equipment, gross
`at`	Total assets
Panel B: Compustat Inputs — Liabilities and equity (9 items)	
`ap`	Accounts payable, trade
`dlc`	Debt in current liabilities
`txp`	Income taxes payable
`lct`	Current liabilities, total
`dltt`	Long-term debt, total
`lt`	Total liabilities
`ceq`	Common/ordinary equity, total
`pstk`	Preferred/preference stock (capital, total)
`re`	Retained earnings
Panel C: Compustat Inputs — Income statement (6 items)	
`sale`	Sales/turnover, net
`cogs`	Cost of goods sold
`dp`	Depreciation and amortization
`txt`	Income taxes, total
`ib`	Income before extraordinary items
`ni`	Net income (loss)
Panel D: Compustat Inputs — Cash flow and market (2 items)	
`sstk`	Sale of common and preferred stock
`csho`	Common shares outstanding
Panel E: Constructed Variables and Indicators Used in Empirical Tables	
Leverage	(`dlc` + `dltt`) / `at`
ROA	`ni` / `at`
Tobin's Q	(CRSP December market equity + `lt`) / `at`
Momentum	Cumulative CRSP stock return over the $[t-12,\ t-2]$-month window (skipping the most recent month), measured at the fiscal year-end month
Book-to-Market	Book equity (Fama–French construction: `seq` + `txditc` − preferred stock, positive values only) / one-month-lagged market capitalization at the fiscal year-end month
Size decile	Within-fiscal-year decile of log(`at`)
Industry	Fama–French 48-industry classification (assigned from SIC)
Top 10% (High Risk)	$\mathbf{1}\{$within-year rank of predicted fraud probability $\geq 0.90\}$, computed separately under each scoring method (Soft Label, Hard Label, RUSBoost)
Within-year percentile rank	Empirical-CDF rank of predicted fraud probability within fiscal year, scaled to $[0, 100]$
Issue$(t+k)$	$\mathbf{1}\{$firm completes an SDC-recorded seasoned equity offering (main tranche, non-withdrawn; IPOs and convertibles excluded) during its fiscal year $t+k\}$

Continued on next page

Table 1 – *Continued from previous page*

Variable	Definition
Issue(t+1..t+3)	**1**{firm issues equity in any of fiscal years $\{t+1, t+2, t+3\}$}
Lag Equity Issue	**1**{firm issued equity in fiscal year $t-1$}
AA Client	**1**{Compustat `au`=1 (Arthur Andersen) in fiscal 2001}, following Dyck et al. (2024)
has_loan	**1**{firm has any active syndicated DealScan facility (start date $\leq$ FY-end t, maturity $\geq$ FY-start t)}, linked to Compustat *gvkey* via the Chava–Roberts DealScan–Compustat link (2024-01 vintage)
FF25 adjusted return	Firm's annual raw stock return − return on the Fama–French 25 (size × book-to-market) benchmark portfolio to which the firm belongs

Table 2: Sample Selection and Annual Distribution

Panel A: the Dechow et al. (2011) AAER filters (upper block; firms counted by company record) and the prediction-sample requirements (lower block; firms counted by distinct GVKEY—the first row collapses records to unique GVKEY firm-years: nine duplicate firm-year records, several company records sharing a GVKEY; the consecutive-years requirement drops no further events). Panel B: five-year sample windows per year, from the training file used by all models. Step 1: complete, consecutive five-year history of the 23 inputs (Table 1). Step 2: windows with soft-label coverage (Section 3.1; CRSP NYSE/AMEX/NASDAQ common equity, share codes 10/11), yielding the 84,089-firm-year analysis sample (783 AAER, all retained). The 2000–2019 rows (54,115; 528 AAER) are the test panel of Table 3; training copies excluded.

Panel A: Fraud-Event Sample Selection

	Misstatement Firm-Years		Firms	
Selection Criterion	Remaining	Dropped	Remaining	Dropped
AAER database filters				
Annual file (earnings misstatement events)	2,195	—	942	—
Less: understatement cases	2,139	56	931	11
Less: missing GVKEY	1,755	384	744	187
Less: no Compustat match or missing total assets	1,619	136	702	42
Less: financial firms (SIC 6000–6999)	1,407	212	609	93
Prediction sample filters				
Less: collapse to unique GVKEY firm-years	1,398	9	597	12
Less: fiscal year outside 1990–2019 analysis window	1,193	205	492	105
Less: fewer than 5 years of prior data	1,012	181	407	85
Less: missing key financial variables	783	229	305	102
Final prediction sample	783		305	

Panel B: Annual Distribution of Five-Year Sample Windows

Year	Step 1 Five-Year Window	Step 2 Analysis Sample	**AAER**
1990	4,023	2,401	9
1991	4,138	2,538	18
1992	4,205	2,657	20
1993	4,300	2,784	14
1994	4,405	2,884	13
1995	4,493	3,043	18
1996	4,704	3,280	24
1997	4,845	3,354	28
1998	4,948	3,409	42
1999	5,301	3,624	69
2000	5,191	3,556	74
2001	4,970	3,380	76
2002	5,148	3,475	64
2003	5,212	3,307	61
2004	5,206	3,179	46
2005	5,027	3,023	37
2006	4,863	2,902	22
2007	4,621	2,742	20
2008	4,441	2,667	14
2009	4,427	2,632	13
2010	4,349	2,548	18
2011	4,235	2,462	15
2012	4,159	2,438	18
2013	4,095	2,371	11
2014	4,069	2,316	14
2015	4,045	2,270	8
2016	4,236	2,269	9
2017	4,180	2,222	5
2018	4,092	2,190	2
2019	3,969	2,166	1
Total	135,897	84,089	783

Table 3: Detection Results by Year, 2000–2019: Soft Label vs. Hard Label vs. RUSBoost vs. Dechow F-score

Out-of-sample detection results for 2000–2019 on the aligned (Soft∩Hard∩RUS) sample, flagging the top 1% of firms within each year by predicted fraud probability (each model's within-year ranking averaged across its 30 training runs). Recall (detection rate) = TP/(TP+FN); precision = TP/flagged; because every method flags the same number of firms each year, the two move together. Total rows report pooled FN/TP counts, the year-by-year mean recall (panel years with no AAER firm-years—2018–2019 in Panel A—enter this mean as zero), and pooled precision. Panel A restricts to firms with total assets > \$750M; Panel B uses the full aligned sample. The Dechow et al. (2011) F-score (Model 1) is re-estimated out of sample under the same expanding-window protocol and ranked within year where its inputs are non-missing.

Year	Soft Label		Hard Label		RUSBoost		F-score		Detection Rate (Recall)				Precision			
	FN	TP	FN	TP	FN	TP	FN	TP	Soft	Hard	RUS	F	Soft	Hard	RUS	F
Panel A: Firms with total assets > \$750M																
2000	35	5	34	6	39	1	39	1	0.125	0.150	0.025	0.025	0.556	0.667	0.111	0.111
2001	41	4	40	5	43	2	42	3	0.089	0.111	0.044	0.067	0.444	0.556	0.222	0.333
2002	28	6	27	7	32	2	33	1	0.176	0.206	0.059	0.029	0.667	0.778	0.222	0.111
2003	23	6	23	6	25	4	29	0	0.207	0.207	0.138	0.000	0.600	0.600	0.400	0.000
2004	18	6	17	7	23	1	24	0	0.250	0.292	0.042	0.000	0.545	0.636	0.091	0.000
2005	13	6	14	5	18	1	19	0	0.316	0.263	0.053	0.000	0.545	0.455	0.091	0.000
2006	6	2	5	3	8	0	8	0	0.250	0.375	0.000	0.000	0.182	0.273	0.000	0.000
2007	7	1	7	1	8	0	8	0	0.125	0.125	0.000	0.000	0.091	0.091	0.000	0.000
2008	5	2	5	2	6	1	7	0	0.286	0.286	0.143	0.000	0.182	0.182	0.091	0.000
2009	5	2	5	2	7	0	7	0	0.286	0.286	0.000	0.000	0.182	0.182	0.000	0.000
2010	6	2	5	3	7	1	7	1	0.250	0.375	0.125	0.125	0.182	0.273	0.091	0.091
2011	4	4	4	4	8	0	8	0	0.500	0.500	0.000	0.000	0.364	0.364	0.000	0.000
2012	9	2	9	2	11	0	10	1	0.182	0.182	0.000	0.091	0.167	0.167	0.000	0.083
2013	7	1	6	2	8	0	8	0	0.125	0.250	0.000	0.000	0.083	0.167	0.000	0.000
2014	6	1	6	1	7	0	6	1	0.143	0.143	0.000	0.143	0.083	0.083	0.000	0.083
2015	2	1	2	1	3	0	3	0	0.333	0.333	0.000	0.000	0.091	0.091	0.000	0.000
2016	4	1	4	1	5	0	5	0	0.200	0.200	0.000	0.000	0.083	0.083	0.000	0.000
2017	1	0	1	0	1	0	1	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2018	0	0	0	0	0	0	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2019	0	0	0	0	0	0	0	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total	220	52	214	58	259	13	264	8	0.192	0.214	0.031	0.024	0.236	0.264	0.059	0.036
Panel B: Full aligned sample																
2000	65	9	66	8	71	3	71	3	0.122	0.108	0.041	0.041	0.250	0.222	0.083	0.083
2001	66	10	65	11	71	5	73	3	0.132	0.145	0.066	0.039	0.294	0.324	0.147	0.088
2002	52	12	51	13	58	6	61	3	0.188	0.203	0.094	0.047	0.343	0.371	0.171	0.086
2003	49	12	48	13	53	8	61	0	0.197	0.213	0.131	0.000	0.353	0.382	0.235	0.000
2004	36	10	36	10	40	6	45	1	0.217	0.217	0.130	0.022	0.312	0.312	0.188	0.031
2005	28	9	28	9	31	6	36	1	0.243	0.243	0.162	0.027	0.290	0.290	0.194	0.032
2006	17	5	18	4	22	0	22	0	0.227	0.182	0.000	0.000	0.167	0.133	0.000	0.000
2007	18	2	18	2	20	0	18	2	0.100	0.100	0.000	0.100	0.071	0.071	0.000	0.071
2008	11	3	11	3	13	1	13	1	0.214	0.214	0.071	0.071	0.111	0.111	0.037	0.037
2009	9	4	9	4	12	1	13	0	0.308	0.308	0.077	0.000	0.148	0.148	0.037	0.000
2010	13	5	12	6	14	4	18	0	0.278	0.333	0.222	0.000	0.192	0.231	0.154	0.000
2011	10	5	10	5	14	1	14	1	0.333	0.333	0.067	0.067	0.200	0.200	0.040	0.040
2012	16	2	16	2	17	1	17	1	0.111	0.111	0.056	0.056	0.080	0.080	0.040	0.040
2013	10	1	9	2	11	0	11	0	0.091	0.182	0.000	0.000	0.042	0.083	0.000	0.000
2014	12	2	12	2	14	0	11	3	0.143	0.143	0.000	0.214	0.083	0.083	0.000	0.125
2015	7	1	7	1	8	0	6	2	0.125	0.125	0.000	0.250	0.043	0.043	0.000	0.087
2016	7	2	7	2	8	1	8	1	0.222	0.222	0.111	0.111	0.087	0.087	0.043	0.043
2017	4	1	4	1	5	0	5	0	0.200	0.200	0.000	0.000	0.043	0.043	0.000	0.000
2018	2	0	2	0	2	0	1	1	0.000	0.000	0.000	0.500	0.000	0.000	0.000	0.045
2019	1	0	1	0	1	0	1	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total	433	95	430	98	485	43	505	23	0.173	0.179	0.061	0.077	0.172	0.178	0.078	0.042

Table 4: Validation Against Post-Scrutiny Arthur Andersen Clients: Soft Label vs. Hard Label vs. RUSBoost vs. Dechow F-score

An out-of-sample *specificity test* on a post-scrutiny anchor: the Arthur Andersen client firm-years of 2002–2005 (2001 AA audit signature, following Dyck et al., 2024), which emerged from the post-Enron re-audits with no AAER. These firms are not certified fraud-free; the scrutiny makes them *relatively* clean, which is all the test requires. Both panels pool 2002–2005, exclude AAER cases, and restrict to firms with total assets > \$750M (the Dyck–Morse–Zingales sample), so the comparison is like-for-like on size; ranks and top-decile flags are computed within this pool, separately for each fiscal year and each training run. Panel A: where the AA firm-years sit in each model's ranking. Panel B: fractional logit (quasi-ML) of the flag probability (the share of the 30 training runs flagging the firm-year in the top decile; for the deterministic F-score, its realized flag, so the estimator reduces to a standard logit) on the AA-client indicator, with the controls shown, year and size-decile×industry fixed effects, and firm-clustered SEs; coefficients are log-odds, with z-statistics in parentheses. Fixed-effect cells in which the flag probability never varies are perfectly separated in the likelihood and are dropped, so N differs across models. The Soft Label estimate (average marginal effect −0.034, i.e. 3.4 percentage points against the 10% baseline flag rate) is corroborated by randomization inference: 300 firm-level reassignments of the AA label within industry×size strata, refitting the model on each draw, give a one-sided $p = 0.017$. "Equality vs. Soft": p-value of a stacked test that the model's AA coefficient equals the Soft Label's. $^{***}/^{**}/^{*}$: 1%/5%/10%.

	Soft Label	Hard Label	RUSBoost	F-score
Panel A: Distribution of Post-Scrutiny AA Firms across Probability Rankings				
In top 10%	7.9%	7.9%	11.1%	9.7%
# AA firm-years in top 10%	34 / 431	34 / 431	48 / 431	42 / 431
Panel B: Specificity test — fractional logit of the flag probability on the AA-client indicator (log-odds)				
AA-Client indicator	−0.427**	−0.352	+0.159	−0.317
	(−2.07)	(−1.57)	(0.57)	(−0.89)
Leverage	−0.101	+0.600	+0.007	+0.991
	(−0.21)	(1.22)	(0.01)	(1.11)
ROA	−1.213***	+1.369**	−3.968***	+1.415
	(−4.00)	(1.97)	(−3.48)	(1.19)
Tobin's Q	−0.160**	−0.299***	+0.041	−0.609***
	(−2.29)	(−2.83)	(0.45)	(−3.98)
Momentum	+0.013	+0.045	+0.673***	+0.089
	(0.12)	(0.43)	(3.83)	(0.48)
Book-to-Market	−0.072	−0.120	−0.182	−0.310
	(−0.51)	(−0.74)	(−0.63)	(−1.45)
Equality vs. Soft (stacked, p)	—	0.49	0.07	0.77
Year FE	Yes	Yes	Yes	Yes
Size-decile × industry FE	Yes	Yes	Yes	Yes
N	2,492	2,481	1,670	1,541
# AA firm-years	401	394	269	236

Table 5: Cross-Sectional Predicted Fraud Probability by Bank-Loan Status: Soft Label vs. Hard Label vs. RUSBoost vs. Dechow F-score

Predicted fraud probability by bank-loan status, 2000–2019, non-AAER firms. *has_loan*=1 if the firm has an active syndicated DealScan facility in fiscal year t (Chava–Roberts link). Panel A reports the fraction of firm-years flagged in the within-year top decile, by loan status (all loan vs. no-loan differences significant at 1%). Panel B reports logistic regressions (log-odds coefficients) of the top-decile flag on *has_loan* with the Graham et al. (2008) bank-loan determinants and cash holdings (committed credit lines substitute for cash; Sufi, 2009), year and size-decile×industry fixed effects, and firm-clustered SEs; fixed-effect cells in which the flag never varies are perfectly separated in the likelihood and are dropped, so N differs across columns. The Altman Z-score control is winsorized at 1/99 and standardized, so its coefficient is per one standard deviation. The F-score is the out-of-sample Dechow et al. (2011) Model 1. ***, **, *: 1%, 5%, 10%.

Panel A: Fraction flagged in within-year top 10%, by loan status (has_loan=1: $N = 28,109$; has_loan=0: $N = 25,478$)

	has_loan=1	has_loan=0
Soft Label	5.5%	15.0%
Hard Label	11.7%	8.2%
RUSBoost	12.6%	7.2%
F-score	12.4%	7.3%

	Soft Label (1)	Hard Label (2)	RUSBoost (3)	F-score (4)
Panel B (logit): DV = $\mathbf{1}\{$firm-year in top-10% of within-year predicted fraud probability$\}$				
has_loan	−0.268***	−0.096	−0.100	−0.074
	(−3.77)	(−1.43)	(−1.35)	(−1.14)
Leverage	+0.267	−0.059	+0.667***	−0.266
	(1.43)	(−0.29)	(2.86)	(−1.34)
ROA	−1.281***	+0.297	−0.241**	+0.497
	(−8.12)	(1.30)	(−2.03)	(0.97)
Tobin's Q	+0.074***	−0.012	+0.106***	+0.194***
	(4.46)	(−0.48)	(5.29)	(5.70)
Tangibility	−2.066***	−2.054***	−6.178***	−21.643***
	(−6.48)	(−7.61)	(−14.30)	(−37.46)
Z-score	−0.481***	−0.199***	−0.163***	+0.227***
	(−10.20)	(−2.71)	(−2.78)	(2.69)
CF volatility	+0.079*	+0.027	+0.019	+0.262***
	(1.91)	(0.64)	(0.53)	(4.16)
Cash holdings	+2.426***	−1.111***	−0.043	−19.749***
	(16.11)	(−5.60)	(−0.21)	(−39.17)
Pseudo R^2 (Panel B)	0.263	0.087	0.270	0.528
Year FE	Yes	Yes	Yes	Yes
Size-decile × industry FE	Yes	Yes	Yes	Yes
Cluster SE (firm)	Yes	Yes	Yes	Yes
N	36,370	39,253	34,773	37,863

Table 6: Equity Issuance and Top-Decile Fraud Probability: Soft Label vs. Hard Label vs. RUSBoost vs. Dechow F-score

Panel A compares SDC seasoned-equity-offering rates (IPOs and convertibles excluded) between each model's top decile and bottom 90% on the non-AAER sample (t-statistics from two-sample tests). Panel B reports logistic regressions (log-odds coefficients) of Issue(t+1..t+3) on the top-decile *High Risk* indicator, controlling for lagged issuance, firm characteristics, and the DeAngelo et al. (2010) SEO determinants (including the Frank and Goyal, 2003 financing deficit), with year and size-decile×industry fixed effects and firm-clustered SEs; fixed-effect cells in which the DV never varies are perfectly separated in the likelihood and are dropped. The R&D control is winsorized at 1/99 and standardized, so its coefficient is per one standard deviation. The F-score is the out-of-sample Dechow et al. (2011) Model 1. ***, **, *: 1%, 5%, 10%.

Panel A: Equity Issuance Rates among Top-Decile Firms

	Soft Label		Hard Label		RUSBoost		F-score		t-statistic (Top–Bot)			
	Top 10%	Bot 90%	Top 10%	Bot 90%	Top 10%	Bot 90%	Top 10%	Bot 90%	Soft	Hard	RUS	F
Issue(t + 1)	15.7%	4.1%	5.1%	5.3%	4.6%	5.3%	6.1%	4.9%	+22.97***	−0.54	−2.26**	+3.30***
Issue(t + 2)	14.1%	3.8%	4.9%	4.9%	3.2%	5.1%	4.7%	4.7%	+21.25***	+0.12	−7.19***	+0.01
Issue(t + 3)	12.4%	3.5%	4.2%	4.4%	3.1%	4.5%	4.2%	4.2%	+19.59***	−0.69	−5.51***	+0.01
Issue(t+1..t+3)	26.8%	9.1%	11.0%	10.9%	9.1%	11.1%	12.1%	10.4%	+28.63***	+0.27	−4.71***	+3.62***
N (firm-years)	5,367	48,220	5,367	48,220	5,367	48,220	5,338	47,932				

Panel B: Logit (log-odds), DV = Issue(t+1..t+3), IV = High Risk (Top Decile of Predicted Probability)

	Soft Label (1)	Hard Label (2)	RUSBoost (3)	F-score (4)
High Risk	+0.400***	+0.248***	+0.058	+0.417***
	(5.77)	(3.16)	(0.71)	(5.64)
Lag Equity Issue	+0.924***	+0.950***	+0.955***	+0.942***
	(16.10)	(16.64)	(16.72)	(16.52)
Leverage	+1.021***	+1.069***	+1.057***	+1.064***
	(6.49)	(6.81)	(6.76)	(6.79)
ROA	−0.219***	−0.297***	−0.294***	−0.329***
	(−2.77)	(−3.50)	(−3.47)	(−3.77)
Tobin's Q	+0.023*	+0.022*	+0.022*	+0.023*
	(1.93)	(1.82)	(1.77)	(1.85)
Momentum	+0.172***	+0.173***	+0.173***	+0.167***
	(8.16)	(8.14)	(8.16)	(7.85)
Book-to-Market	−0.055	−0.060	−0.059	−0.053
	(−1.31)	(−1.42)	(−1.39)	(−1.25)
Cash/Assets	−0.094	+0.051	+0.034	+0.233
	(−0.65)	(0.35)	(0.23)	(1.58)
Firm Age	−0.021***	−0.021***	−0.021***	−0.021***
	(−6.96)	(−6.99)	(−7.03)	(−6.98)
R&D	+0.094***	+0.096***	+0.096***	+0.092***

Continued on next page

Table 6 — Continued from previous page

Panel B: Logit (log-odds), DV = Issue(t+1..t+3), IV = High Risk (Top Decile of Predicted Probability)

	Soft Label (1)	Hard Label (2)	RUSBoost (3)	F-score (4)
	(4.84)	(5.01)	(4.96)	(4.78)
Idio. Volatility	−0.075	−0.061	−0.063	−0.063
	(−1.34)	(−1.11)	(−1.15)	(−1.15)
Financing Deficit	+0.719***	+0.764***	+0.764***	+0.695***
	(6.75)	(7.01)	(7.01)	(6.42)
Dividend Payer	−0.650***	−0.652***	−0.652***	−0.636***
	(−7.54)	(−7.60)	(−7.56)	(−7.39)
Year FE	Yes	Yes	Yes	Yes
Size-decile × Industry FE	Yes	Yes	Yes	Yes
N	37,850	37,850	37,850	37,849
Pseudo R^2	0.182	0.181	0.180	0.182

Table 7: Post-Issuance FF25 Adjusted Stock Returns: Soft Label vs. Hard Label vs. RUSBoost vs. Dechow F-score

Winsorized (1st/99th pct.) mean annual FF25-adjusted returns for equity issuers in each model's top decile. FF25 adjusted return = firm's annual raw return − matched Fama–French 25 (size×BM) portfolio return. Panel A uses value-weighted (VW) benchmark; Panel B uses equal-weighted (EW). "Soft-only" ("Hard-only"): top decile under Soft (Hard) only. Avg[1:5]: per-firm mean over the $[t+1, t+5]$ window. Stars from one-sample t-tests vs. zero. AAER firms excluded. ***, **, *: 1%, 5%, 10%.

	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	Avg[1:5]	N
Panel A: Value-Weighted (VW) FF25 Adjusted Returns							
Soft top-decile issuers	−10.9%***	−14.6%***	−12.2%***	−10.7%***	−4.0%	−11.4%***	961
Soft-only (non-overlap)	−13.0%***	−15.0%***	−12.7%***	−13.2%***	−5.6%	−12.8%***	859
Hard top-decile issuers	−3.1%	−3.0%	−6.2%**	−3.1%	+1.1%	−3.5%**	346
Hard-only (non-overlap)	−6.9%**	+0.3%	−5.4%*	−7.4%**	−1.1%	−5.4%***	244
RUSBoost top-decile issuers	−3.5%*	−0.6%	+1.8%	+1.9%	−0.2%	+0.3%	546
F-score top-decile issuers	−2.9%	−7.3%***	−3.9%	−3.8%	−1.0%	−3.5%**	383
Full sample issuers	−6.1%***	−7.6%***	−4.8%***	−6.8%***	−4.1%***	−6.6%***	3,115
Full sample non-issuers	+1.4%***	+1.5%***	+0.6%**	+0.5%**	+0.2%	+0.7%***	37,564
Panel B: Equal-Weighted (EW) FF25 Adjusted Returns							
Soft top-decile issuers	−12.0%***	−14.5%***	−11.4%***	−9.3%***	−3.1%	−11.1%***	961
Soft-only (non-overlap)	−14.0%***	−14.8%***	−11.8%***	−11.6%***	−4.5%	−12.5%***	859
Hard top-decile issuers	−5.2%*	−3.8%	−6.1%**	−2.7%	+0.8%	−4.1%**	346
Hard-only (non-overlap)	−9.2%***	−0.5%	−5.2%	−6.8%*	−1.7%	−5.9%***	244
RUSBoost top-decile issuers	−4.9%**	−1.5%	+1.5%	+1.9%	−0.2%	−0.4%	546
F-score top-decile issuers	−4.3%	−8.0%***	−4.0%	−3.7%	−1.5%	−4.0%**	383
Full sample issuers	−7.5%***	−8.2%***	−4.8%***	−6.4%***	−3.8%***	−6.9%***	3,115
Full sample non-issuers	−0.4%	+0.6%**	+0.3%	+0.3%	+0.0%	−0.1%	37,564

Table 8: Equal-Weighted Long-Short Portfolio Factor Model Tests: Soft Label vs. Hard Label vs. RUSBoost vs. Dechow F-score

Fama–French factor regressions of monthly long-short portfolio returns (bottom 90% − top 10%; AAER firms excluded; equal-weighted; annual rebalancing with a three-year overlapping holding period; signal years 2000–2019). Portfolios follow the Fama–French timing convention: a signal from the fiscal year ending in calendar year T enters at the end of June of $T+1$, so every constituent's annual report is public before it is traded on (at least a six-month lag for December fiscal-year ends). Constituent monthly returns are delisting-adjusted and winsorized at 1%/99% before portfolio formation. Alphas are annualized and in percent; t-statistics use Newey–West (12 lags) SEs. The F-score is the out-of-sample Dechow et al. (2011) Model 1. ***, **, *: 1%, 5%, 10%.

	α (%)	β_{MKT}	β_{SMB}	β_{HML}	β_{UMD}	β_{RMW}	β_{CMA}	R^2
Panel A: Soft Label								
3-Factor	13.005***	−0.131	−0.418***	+0.462***				0.314
	(4.86)	(−1.56)	(−3.07)	(6.84)				
4-Factor	12.132***	−0.075	−0.409***	+0.501***	+0.132*			0.340
	(4.78)	(−1.12)	(−3.01)	(6.86)	(1.80)			
5-Factor	7.503***	−0.031	−0.183*	+0.389***		+0.820***	+0.068	0.548
	(3.09)	(−0.51)	(−1.95)	(7.94)		(10.05)	(0.65)	
Panel B: Hard Label								
3-Factor	−0.249	−0.081***	+0.080***	−0.025				0.152
	(−0.33)	(−3.41)	(2.87)	(−0.81)				
4-Factor	−0.472	−0.065***	+0.082***	−0.013	+0.038**			0.177
	(−0.62)	(−3.26)	(3.09)	(−0.43)	(2.30)			
5-Factor	−0.716	−0.072***	+0.094***	−0.072**		+0.042	+0.071*	0.174
	(−0.91)	(−3.38)	(3.00)	(−2.57)		(0.83)	(1.73)	
Panel C: RUSBoost								
3-Factor	−0.345	−0.161***	+0.160***	+0.093*				0.208
	(−0.24)	(−3.43)	(3.13)	(1.83)				
4-Factor	−0.865	−0.123***	+0.166***	+0.119**	+0.088*			0.256
	(−0.66)	(−3.94)	(3.43)	(2.16)	(1.89)			
5-Factor	−1.515	−0.140***	+0.212***	+0.027		+0.182	+0.028	0.260
	(−1.24)	(−4.61)	(3.70)	(0.68)		(1.54)	(0.30)	
Panel D: Dechow F-score (Model 1, OOS)								
3-Factor	0.422	−0.047*	+0.020	+0.045				0.036
	(0.44)	(−1.92)	(0.41)	(1.26)				
4-Factor	0.103	−0.025	+0.024	+0.061*	+0.054**			0.068
	(0.12)	(−1.42)	(0.50)	(1.70)	(2.01)			
5-Factor	0.251	−0.044*	+0.024	+0.027		+0.012	+0.033	0.038
	(0.29)	(−1.81)	(0.57)	(0.81)		(0.12)	(0.47)	

Table 9: Decomposing the Training Target: Enforcement Anchor versus Forward Tilt

Ablation of the soft-label target with the network held fixed: the Hard Label CNN trains on the enforcement anchor alone (binary AAER); the de-anchored CNN trains on the forward tilt alone (AAER firm-years receive their own accounting count label, fraud replication removed, the label-neutralization rule applied to all windows); the Soft Label combines anchor and tilt. Architecture, inputs, hyperparameters (the pinned configuration, no re-selection), expanding windows, and the 30-run ensemble convention are identical across the three networks. Panel A reports yearly-mean AAER recall: the Hard Label and Soft Label are evaluated on the full 2000–2019 aligned sample of Table 3, so their rates match that table; the de-anchored CNN does not score 42 AAER firm-years that lack a forward count label and is evaluated on the remaining firm-years, which leaves its near-zero detection unaffected. Panel B reports each footprint measure as a single differential. *Excess SEO rate of flagged firms*: the three-year (t+1..t+3) SDC seasoned-equity-offering rate of the within-year top decile minus that of the bottom 90% (non-AAER pool; Welch t in parentheses). *Excess flag rate of post-scrutiny AA clients*: the top-decile flag rate of the Arthur Andersen clients of 2002–2005 minus that of other large firms (> \$750M, non-AAER). *Excess flag rate of bank borrowers*: the flag rate of active syndicated-loan holders minus that of non-borrowers (2000–2019, non-AAER). Hidden-fraud capture predicts a positive issuance differential and negative AA-client and bank-loan differentials. Underlying levels for the AA and bank-loan comparisons appear in Panel A of Tables 4 and 5; differentials are computed from unrounded rates and may differ from differences of the rounded levels by 0.1 pp. The size-controlled formal tests are in those tables' Panel B, with the de-anchored CNN's coefficients reported in Section 6.1. Panel C reports top-decile overlap with the Soft flag set on the common sample. A transparent linear ROA-decline forecast (a non-CNN placebo for the tilt), evaluated on the subset with complete characteristics, likewise detects essentially no enforced fraud and shows no issue-and-underperform footprint (untabulated).

	Hard Label (anchor only)	De-anchored CNN (tilt only)	Soft Label (anchor + tilt)
Panel A: Detection of enforced fraud (yearly-mean AAER recall, top-1% cutoff)			
Top-1% detection rate	17.9%	1.3%	17.3%
Panel B: Hidden-fraud footprint (each row a differential)			
Excess SEO rate of flagged firms (pp)	+0.1	+3.0	+17.7
	(0.27)	(6.22)	(28.63)
Excess flag rate of post-scrutiny AA clients (pp)	−2.6	+0.2	−2.6
Excess flag rate of bank borrowers (pp)	+3.6	−1.5	−9.5
Panel C: Relation to the Soft Label ranking			
Top-decile overlap with Soft flags	25.6%	30.3%	—

Table 10: Industry Sentiment Sensitivity and Predicted-Fraud Flag Rates

Industry-year panel, FF48 level, 2000–2019: $FlagRate_{i,t} = \delta\,(SENT_t \times E_i) + \lambda\,E_i \times (t-2000) + \boldsymbol{\theta}'\mathbf{W}_{i,t} + \alpha_i + \tau_t + u_{i,t}$, SEs clustered by industry. *FlagRate*: the percentage of non-AAER firms flagged under the model's fixed in-sample top-decile cutoff, in percentage points (so coefficients are in percentage points; t-statistics are invariant to the scaling); *SENT*: standardized macro-orthogonalized Baker and Wurgler (2006) index. The exposure E_i is the industry share of firms in the top decile of the published SYY mispricing score (Stambaugh and Yuan, 2017) (left block) or of an SYY-style composite of the five Hribar and McInnis (2012) valuation-uncertainty characteristics (right block), z-scored across industries. Controls $\mathbf{W}_{i,t}$: industry means of log size, leverage, book-to-market, retained earnings/assets, asset growth, sales growth, ROA, and the prior fiscal year's stock return, winsorized 1%/99%. The return control is lagged so that it is pre-determined relative to the flag rate it conditions: a return compounded over year t is realized over the same window as the year-t flags and, for non-December fiscal-year ends, extends past the prediction date. The interaction δ is unaffected by this choice. Coefficients on $\mathbf{W}_{i,t}$ are reported for completeness and are not the object of inference. RUSBoost's right-block interaction is marginally positive only without the trend control (untabulated: +0.28, $t = 1.89$) and is not robust to alternative exposures or to excluding pharmaceuticals. Section 6.2 details the construction. $^{*}/^{**}/^{***}$: 10%/5%/1%.

	Dependent variable: industry-year flag rate $FlagRate_{i,t}$ (FF48)							
	E_i: published SYY mispricing score				E_i: SYY composite of HM characteristics			
	Soft	Hard	RUS	F-score	Soft	Hard	RUS	F-score
$SENT_t \times E_i$	+0.921***	+0.326	+0.041	−0.224*	+1.362***	+0.179	+0.134	−0.006
	(4.54)	(1.40)	(0.27)	(−1.65)	(7.17)	(0.92)	(1.08)	(−0.03)
$E_i \times (t-2000)$	−0.088	+0.131**	−0.032	−0.014	−0.180***	+0.245**	−0.068	−0.004
	(−0.93)	(2.26)	(−0.66)	(−0.36)	(−2.98)	(2.40)	(−1.54)	(−0.08)
Mean log size	+0.433	+1.933	+0.036	−0.085	−0.481	+2.710	−0.266	−0.100
	(0.28)	(0.93)	(0.03)	(−0.07)	(−0.33)	(1.26)	(−0.28)	(−0.08)
Leverage	−3.067	+11.373***	−0.120	+2.128	−0.185	+9.730***	+0.568	+2.388
	(−0.58)	(3.03)	(−0.06)	(0.84)	(−0.04)	(2.82)	(0.30)	(0.96)
Book-to-market	−0.291	+3.822**	−0.215	−2.817	−0.640	+4.381***	−0.351	−2.811
	(−0.19)	(2.52)	(−0.19)	(−1.56)	(−0.39)	(2.75)	(−0.30)	(−1.52)
Retained earnings	−0.313	+0.277	−0.046	+0.073	−0.138	+0.181	−0.006	+0.079
	(−1.15)	(1.02)	(−0.34)	(0.45)	(−0.58)	(0.69)	(−0.04)	(0.50)
Asset growth	−1.586	−1.403	+2.068	+3.276	+0.040	−3.057	+2.622	+3.470
	(−0.79)	(−0.61)	(1.30)	(1.51)	(0.02)	(−1.25)	(1.61)	(1.52)
Sales growth	+5.184	−0.562	−0.868	−0.022	+3.465	−0.050	−1.190	−0.083
	(1.58)	(−0.18)	(−0.49)	(−0.01)	(1.25)	(−0.02)	(−0.70)	(−0.03)
ROA	−0.078	+3.541	−0.640	+2.356*	+0.107	+3.989	−0.708	+2.348*
	(−0.03)	(1.40)	(−0.59)	(1.74)	(0.04)	(1.57)	(−0.69)	(1.72)
Prior-year return	+0.656	+0.285	+0.965*	+0.806	+0.511	+0.772	+0.853	+0.659
	(0.45)	(0.13)	(1.76)	(0.84)	(0.36)	(0.34)	(1.52)	(0.69)
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	523	523	523	522	523	523	523	522

Appendix A CNN Architecture, Loss Function, and Hyperparameter Configuration

This appendix provides the implementation details of the CNN architecture and loss function referenced in Section 3.1 and the Model Architecture subsection of the main text. As emphasized there, the architectural choices below are not the conceptual contribution of the paper, and the soft-label idea is in principle compatible with alternative architectures suitable for tabular time-series inputs.

Appendix A.1 Convolutional Network with Residual Connections

Convolutional Neural Networks (CNNs) automatically learn hierarchical feature representations through convolutional operations that capture spatial and temporal patterns (LeCun et al., 2015). We adapt the architecture to financial-statement data with residual connections (He et al., 2016) that facilitate gradient flow during training. The convolution operation applies a set of learnable filters across the input matrix:

$$Y_{i,j} = \sigma\left(\sum_m \sum_n W_{m,n} \cdot X_{i+m,j+n} + b\right), \tag{6}$$

where σ is the activation function and b is the bias. The network learns the filters W through backpropagation.

First convolutional block. The input passes through a convolutional layer with padding to maintain spatial dimensions. The kernel has height one in the year dimension and detects relationships among adjacent financial variables within the same time period (e.g., interactions among revenue, cost of goods sold, and gross profit, or among different expense categories); all cross-year interactions are formed by the fully connected layers, for the reasons discussed in the Model Architecture subsection of the main text. We

apply batch normalization, an Exponential Linear Unit (ELU) activation

$$\mathrm{ELU}(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(e^x - 1) & \text{if } x \leq 0, \end{cases} \tag{7}$$

with $\alpha = 1$, and dropout regularization. Unlike ReLU, ELU allows negative outputs and yields smoother gradients, which helps prevent the "dying ReLU" problem.

Residual connection. A skip connection passes the input through a separate convolutional layer to match dimensions and adds it to the output of the first block:

$$\mathrm{Output}_1 = \mathrm{BN}(\mathrm{ELU}(F_1(X) + G_1(X))), \tag{8}$$

where F_1 is the main convolutional path (Conv $\rightarrow$ BN $\rightarrow$ ELU $\rightarrow$ Dropout), G_1 is the residual path (Conv $\rightarrow$ BN), and BN denotes batch normalization. The skip connection alleviates the vanishing-gradient problem during backpropagation, allowing the network to learn incremental refinements rather than complete transformations.

Second convolutional block. The output of the first block passes through a second convolutional layer, followed by batch normalization, ELU activation, and dropout. This block captures pairwise interactions between the higher-level features produced by the first layer.

Fully connected layers and output. After the convolutional layers, the feature maps are flattened and passed through three fully connected (dense) layers, each with ELU activation and dropout. The final layer produces a scalar passed through a sigmoid activation to yield a fraud probability:

$$\hat{y} = \sigma(W_3 \cdot h_2 + b_3) = \frac{1}{1 + e^{-(W_3 \cdot h_2 + b_3)}}, \tag{9}$$

where h_2 is the output of the previous dense layer. Figure 3 illustrates the complete architecture.

Appendix A.2 Smooth Certainty-Weighted Focal Loss: Full Specification

The loss function in the main text builds on the Focal Loss framework of Lin et al. (2017), which addresses class imbalance by down-weighting well-classified examples and focusing on hard-to-classify cases. The full per-observation loss is

$$\mathcal{L}_i = w(y_i) \cdot \alpha \, (1 - p_t)^{\gamma} \cdot \mathrm{BCE}(y_i, \hat{y}_i), \tag{10}$$

where α is the balancing parameter and γ is the focusing parameter that reduces the relative loss for well-classified examples. The certainty weight $w(y_i)$ is

$$w(y_i) = w_{\min} + (w_{\max} - w_{\min}) \cdot \mathrm{certainty}(y_i), \qquad \mathrm{certainty}(y_i) = 2|y_i - 0.5|, \tag{11}$$

where $w_{\min}$ and $w_{\max}$ are tunable parameters controlling the weight range. The weight rises monotonically from $w_{\min}$ at $y_i = 0.5$ (maximum uncertainty) to $w_{\max}$ at $y_i \in \{0, 1\}$ (hard labels):

- When $y_i = 0$ or $y_i = 1$ (hard labels or very confident soft labels): $w_i = w_{\max}$.
- When $y_i = 0.5$ (maximum uncertainty): $w_i = w_{\min}$.

An advantage of this formulation is its smoothness and differentiability. Unlike threshold-based weighting schemes that can create gradient discontinuities, our symmetric certainty function produces stable gradients throughout training, facilitating effective optimization via backpropagation.

Appendix A.3 Optimization and Dropout

Optimization. We use the Adam optimizer with an adaptive learning rate that adjusts based on first and second moment estimates of gradients. The model is trained using mini-batch gradient descent. All random seeds for weight initialization, data shuffling, and dropout operations are fixed to ensure reproducibility and deterministic training.

Dropout regularization. Following Srivastava et al. (2014), we apply dropout after the first convolutional block, after the first dense layer, and after the second dense layer. Dropout works by randomly deactivating a subset of neurons during each training iteration. This stochastic process means that even when the same observation is presented to the network multiple times (as occurs under our fraud-replication augmentation), the model encounters it through different effective network architectures due to varying dropout patterns. Each replicated observation therefore contributes differently to the learning process, preventing the model from memorizing repeated instances. The network instead learns features that generalize across different subnetwork configurations, transforming simple data replication into genuine augmentation through multiple distinct dropout-induced "views" of each fraud case.

Appendix B Hyperparameter Grids and Configuration Selection

Appendix B.1 The configuration grid

Because neural-network results can be sensitive to training hyperparameters, we do not report a single hand-picked model. We train a grid of ten configurations, and we use the *same* ten configurations for both the Soft Label and the Hard Label networks, so that every Soft–Hard comparison in the paper holds the candidate set fixed. The grid (Table B1) spans the dimensions that matter most for this architecture: training length (60 to 150 epochs), learning rate (1.0 to 4.0×10^{-4}), batch size (128 or 256), network capacity (40/20 convolutional filters with 160/80 dense units, or 48/24 with 192/96), dropout (0.075 to 0.085), and the focal-loss parameters ($\alpha \in [0.35, 0.42]$, $\gamma \in [2.8, 3.2]$). The certainty-weight range of the loss is held fixed at $[w_{\min}, w_{\max}] = [0.2, 5.0]$ throughout. Each configuration is trained as 30 independent runs (different random seeds for initialization, shuffling, and dropout), and all results in the paper use the 30-run-averaged predictions, or the per-run flags where noted (e.g., the flag-probability dependent variable of Table 4, Panel B), so that no result depends on a single training draw. The RUSBoost benchmark is treated symmetrically: we estimate a ten-configuration, 30-runs-per-configuration grid that varies one hyperparameter at a time around the published Bao et al. (2020) baseline (Table B1, Panel B): the number of boosting iterations (the boosting analogue of training length), the learning rate, the minimum leaf size of the base decision trees, and the undersampling ratio.

Appendix B.2 The pre-committed selection rule

The paper's main tables report one pinned configuration per label. The selection was made once, by a rule fixed in advance, using only data that predates the 2000–2019 reporting window. The timeline is: training samples with window-ends through 1997; 1998 skipped by the two-year label-masking gap (Section 3.4.3); fiscal year 1999 as the

validation fold; and 2000–2019 as the disjoint reporting window. We use the single 1999 fold rather than an average over earlier folds because the earlier folds are fraud-sparse (models trained through 1994 see only 74 fraud windows), so the validation signal in those folds is weak; 1999 is the most saturated fold and the closest to the reporting window.

Each label is selected on *its own training objective*. For the Hard Label, the criterion is the 1999 out-of-sample AUC against the AAER indicator, the binary objective the model is trained on. For the Soft Label, the criterion is the Spearman rank correlation between the run-averaged 1999 out-of-sample prediction and the soft label among *non-AAER* firms. This metric follows from the design in three steps: the graded non-AAER signal is the component of the Soft training target that is unique to it (the fraud anchor at 1.0 is shared with the Hard Label, and including AAER firms would turn the criterion into a recall proxy, the enforcement-based yardstick the paper argues is contaminated); the soft label is ordinal by construction (a monotone transform of the indicator count, with an arbitrary scale parameter), so a rank metric is the appropriate fit measure; and every downstream use of the model output is itself rank-based (within-year percentile ranks and top-decile flags). The criterion touches none of the paper's economic outcome variables, so the selection is non-circular with respect to every validation test.

Applying the rule across the ten configurations of Table B1 (the table lists hyperparameters; the validation metrics quoted in this paragraph are untabulated), Panel A selects configuration 3 for the Soft Label ($\rho = 0.232$; it also wins under Kendall's τ, under the all-firm correlation, and in 84% of run-level bootstrap resamples) and configuration 4 for the Hard Label (AUC = 0.709). The RUSBoost benchmark requires no selection: it is pinned to the published Bao et al. (2020) configuration (Panel B, configuration 3: 300 boosting iterations, learning rate 0.1), and its grid serves as a robustness check around that baseline. Two features of the validation results are worth noting. First, the metrics are tightly clustered across configurations (ρ from 0.204 to 0.232; AUC from 0.692 to 0.709), consistent with the robustness of the paper's results across the grid. Second, the same clustering does not hold *across labels*: in the same validation year, the Hard Label

models' predictions are essentially uncorrelated with the graded non-AAER soft signal (rank correlations of 0.00–0.04, against 0.20–0.23 for every Soft configuration), an early validation-stage indication that the two training targets produce genuinely different rankings of the non-AAER pool. The de-anchored network of Section 6.1 deliberately inherits the Soft Label's pinned configuration without re-selection: the ablation's logic requires that architecture, hyperparameters, and training protocol be held fixed so that the supervisory target is the only difference.

Table B1: The Hyperparameter Configuration Grids

Panel A: the ten CNN training configurations estimated for both the Soft Label and Hard Label networks, ordered by training length. Conv C_1/C_2: filters in the two convolutional blocks; Dense D_1/D_2: units in the first two fully connected layers; LR in units of 10^{-4}; the same dropout rate is applied at the three dropout sites; α and γ are the focal-loss parameters (certainty-weight range fixed at $[0.2, 5.0]$). The pre-committed validation rule described above selects configuration 3 for the Soft Label and configuration 4 for the Hard Label. Panel B: the ten RUSBoost configurations, ordered by the number of boosting iterations, varying one hyperparameter at a time around the published Bao et al. (2020) baseline (configuration 3, the pinned benchmark): the number of boosting iterations, the learning rate, the minimum leaf size of the base decision trees, and the undersampling ratio (majority-to-minority ratio after random undersampling; 1.0 = balanced classes). Every configuration in both panels is trained as 30 independent runs, and all reported results use run-averaged predictions.

Panel A: CNN configurations (Soft Label and Hard Label)

ID	Epochs	Batch	LR (10^{-4})	Conv C_1/C_2	Dense D_1/D_2	Dropout	α	γ
1	60	128	3.5	40/20	160/80	0.075	0.35	2.8
2	60	128	3.5	40/20	160/80	0.075	0.40	2.8
3	60	128	3.5	48/24	192/96	0.080	0.35	3.0
4	80	128	3.0	40/20	160/80	0.075	0.40	2.8
5	80	128	3.0	48/24	192/96	0.080	0.35	3.0
6	90	128	2.0	40/20	160/80	0.080	0.42	3.2
7	100	128	1.0	40/20	160/80	0.080	0.35	2.8
8	100	256	4.0	40/20	160/80	0.075	0.38	3.0
9	120	128	1.0	40/20	160/80	0.085	0.38	3.0
10	150	128	1.5	40/20	160/80	0.085	0.38	3.0

Panel B: RUSBoost configurations

ID	Boosting iterations	Learning rate	Min. leaf size	Undersampling ratio
1	100	0.10	5	1.0
2	200	0.10	5	1.0
3	300	0.10	5	1.0
4	300	0.05	5	1.0
5	300	0.20	5	1.0
6	300	0.10	10	1.0
7	300	0.10	20	1.0
8	300	0.10	5	0.5
9	500	0.10	5	1.0
10	1,000	0.10	5	1.0

Appendix C Detection Results over Alternative Sample Windows

This appendix reproduces the by-year detection table (Table 3; firms with total assets > \$750M in Panel A, full aligned sample in Panel B) for the shorter out-of-sample sub-windows 2000–2008 (Table C1) and 2000–2014 (Table C2).

Table C1: Detection Results by Year, 2000–2008: Soft Label vs. Hard Label vs. RUSBoost vs. Dechow F-score

Out-of-sample detection results for 2000–2008 on the aligned sample, reproducing Table 3 for the shorter sub-window; definitions and construction as there. Panel A: firms with total assets > \$750M; Panel B: full aligned sample.

Year	Soft Label		Hard Label		RUSBoost		F-score		Detection Rate (Recall)				Precision			
	FN	TP	FN	TP	FN	TP	FN	TP	Soft	Hard	RUS	F	Soft	Hard	RUS	F
Panel A: Firms with total assets > \$750M																
2000	35	5	34	6	39	1	39	1	0.125	0.150	0.025	0.025	0.556	0.667	0.111	0.111
2001	41	4	40	5	43	2	42	3	0.089	0.111	0.044	0.067	0.444	0.556	0.222	0.333
2002	28	6	27	7	32	2	33	1	0.176	0.206	0.059	0.029	0.667	0.778	0.222	0.111
2003	23	6	23	6	25	4	29	0	0.207	0.207	0.138	0.000	0.600	0.600	0.400	0.000
2004	18	6	17	7	23	1	24	0	0.250	0.292	0.042	0.000	0.545	0.636	0.091	0.000
2005	13	6	14	5	18	1	19	0	0.316	0.263	0.053	0.000	0.545	0.455	0.091	0.000
2006	6	2	5	3	8	0	8	0	0.250	0.375	0.000	0.000	0.182	0.273	0.000	0.000
2007	7	1	7	1	8	0	8	0	0.125	0.125	0.000	0.000	0.091	0.091	0.000	0.000
2008	5	2	5	2	6	1	7	0	0.286	0.286	0.143	0.000	0.182	0.182	0.091	0.000
Total	176	38	172	42	202	12	209	5	0.203	0.224	0.056	0.013	0.413	0.457	0.130	0.054
Panel B: Full aligned sample																
2000	65	9	66	8	71	3	71	3	0.122	0.108	0.041	0.041	0.250	0.222	0.083	0.083
2001	66	10	65	11	71	5	73	3	0.132	0.145	0.066	0.039	0.294	0.324	0.147	0.088
2002	52	12	51	13	58	6	61	3	0.188	0.203	0.094	0.047	0.343	0.371	0.171	0.086
2003	49	12	48	13	53	8	61	0	0.197	0.213	0.131	0.000	0.353	0.382	0.235	0.000
2004	36	10	36	10	40	6	45	1	0.217	0.217	0.130	0.022	0.312	0.312	0.188	0.031
2005	28	9	28	9	31	6	36	1	0.243	0.243	0.162	0.027	0.290	0.290	0.194	0.032
2006	17	5	18	4	22	0	22	0	0.227	0.182	0.000	0.000	0.167	0.133	0.000	0.000
2007	18	2	18	2	20	0	18	2	0.100	0.100	0.000	0.100	0.071	0.071	0.000	0.071
2008	11	3	11	3	13	1	13	1	0.214	0.214	0.071	0.071	0.111	0.111	0.037	0.037
Total	342	72	341	73	379	35	400	14	0.182	0.181	0.077	0.039	0.251	0.254	0.122	0.049

Table C2: Detection Results by Year, 2000–2014: Soft Label vs. Hard Label vs. RUSBoost vs. Dechow F-score

Out-of-sample detection results for 2000–2014 on the aligned sample, reproducing Table 3 for the shorter sub-window; definitions and construction as there. Panel A: firms with total assets > \$750M; Panel B: full aligned sample.

Year	Soft Label		Hard Label		RUSBoost		F-score		Detection Rate (Recall)				Precision			
	FN	TP	FN	TP	FN	TP	FN	TP	Soft	Hard	RUS	F	Soft	Hard	RUS	F
Panel A: Firms with total assets > \$750M																
2000	35	5	34	6	39	1	39	1	0.125	0.150	0.025	0.025	0.556	0.667	0.111	0.111
2001	41	4	40	5	43	2	42	3	0.089	0.111	0.044	0.067	0.444	0.556	0.222	0.333
2002	28	6	27	7	32	2	33	1	0.176	0.206	0.059	0.029	0.667	0.778	0.222	0.111
2003	23	6	23	6	25	4	29	0	0.207	0.207	0.138	0.000	0.600	0.600	0.400	0.000
2004	18	6	17	7	23	1	24	0	0.250	0.292	0.042	0.000	0.545	0.636	0.091	0.000
2005	13	6	14	5	18	1	19	0	0.316	0.263	0.053	0.000	0.545	0.455	0.091	0.000
2006	6	2	5	3	8	0	8	0	0.250	0.375	0.000	0.000	0.182	0.273	0.000	0.000
2007	7	1	7	1	8	0	8	0	0.125	0.125	0.000	0.000	0.091	0.091	0.000	0.000
2008	5	2	5	2	6	1	7	0	0.286	0.286	0.143	0.000	0.182	0.182	0.091	0.000
2009	5	2	5	2	7	0	7	0	0.286	0.286	0.000	0.000	0.182	0.182	0.000	0.000
2010	6	2	5	3	7	1	7	1	0.250	0.375	0.125	0.125	0.182	0.273	0.091	0.091
2011	4	4	4	4	8	0	8	0	0.500	0.500	0.000	0.000	0.364	0.364	0.000	0.000
2012	9	2	9	2	11	0	10	1	0.182	0.182	0.000	0.091	0.167	0.167	0.000	0.083
2013	7	1	6	2	8	0	8	0	0.125	0.250	0.000	0.000	0.083	0.167	0.000	0.000
2014	6	1	6	1	7	0	6	1	0.143	0.143	0.000	0.143	0.083	0.083	0.000	0.083
Total	213	50	207	56	250	13	255	8	0.221	0.250	0.042	0.032	0.311	0.348	0.081	0.050
Panel B: Full aligned sample																
2000	65	9	66	8	71	3	71	3	0.122	0.108	0.041	0.041	0.250	0.222	0.083	0.083
2001	66	10	65	11	71	5	73	3	0.132	0.145	0.066	0.039	0.294	0.324	0.147	0.088
2002	52	12	51	13	58	6	61	3	0.188	0.203	0.094	0.047	0.343	0.371	0.171	0.086
2003	49	12	48	13	53	8	61	0	0.197	0.213	0.131	0.000	0.353	0.382	0.235	0.000
2004	36	10	36	10	40	6	45	1	0.217	0.217	0.130	0.022	0.312	0.312	0.188	0.031
2005	28	9	28	9	31	6	36	1	0.243	0.243	0.162	0.027	0.290	0.290	0.194	0.032
2006	17	5	18	4	22	0	22	0	0.227	0.182	0.000	0.000	0.167	0.133	0.000	0.000
2007	18	2	18	2	20	0	18	2	0.100	0.100	0.000	0.100	0.071	0.071	0.000	0.071
2008	11	3	11	3	13	1	13	1	0.214	0.214	0.071	0.071	0.111	0.111	0.037	0.037
2009	9	4	9	4	12	1	13	0	0.308	0.308	0.077	0.000	0.148	0.148	0.037	0.000
2010	13	5	12	6	14	4	18	0	0.278	0.333	0.222	0.000	0.192	0.231	0.154	0.000
2011	10	5	10	5	14	1	14	1	0.333	0.333	0.067	0.067	0.200	0.200	0.040	0.040
2012	16	2	16	2	17	1	17	1	0.111	0.111	0.056	0.056	0.080	0.080	0.040	0.040
2013	10	1	9	2	11	0	11	0	0.091	0.182	0.000	0.000	0.042	0.083	0.000	0.000
2014	12	2	12	2	14	0	11	3	0.143	0.143	0.000	0.214	0.083	0.083	0.000	0.125
Total	412	91	409	94	461	42	484	19	0.194	0.202	0.074	0.046	0.208	0.215	0.096	0.044